Temporal Features as Measures of Tie Strength in Mobile Phone Networks

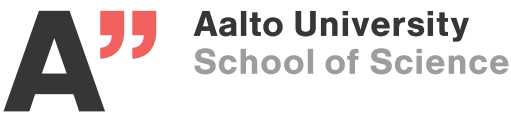
Javier Ureña Carrión

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.
Espoo 27.5.2019

Thesis supervisor and advisor:

Prof. Mikko Kivelä



AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Javier Ureña Carrión

Title: Temporal Features as Measures of Tie Strength in Mobile Phone Networks

Date: 27.5.2019 Language: English Number of pages: 6+83

Department of Computer Science

Professorship: Complex Systems

Supervisor and advisor: Prof. Mikko Kivelä

The use of auto-recorded communication data, such as mobile phone call logs, has reshaped our capacity to model and understand of social systems. In such studies, the strength of a tie between two people has been of great value from both theoretical and sociological perspectives, yet it is not easy to quantify. Tie strengths are commonly measured in terms of communication intensity (number or duration of calls, etc) as a form of convenience rather than a justified choice, yet these intensity-based measures do not uncover the myriad of ways in which such intensity takes place, hindering information about the strength of ties.

Here, we conceive tie strength as a latent variable we want to predict based on features of the time sequences of interactions. We assume that tie strength is expressed as the structural overlap in social networks, in a manner inspired by Granovetter's hypothesis, where strong ties are embedded in community structures, while weak ties serve as inter-community bridges. With this assumption, we use temporal and static features to predict overlap in lieu of the latent tie strength. We analyze a mobile phone dataset of 6.5 million people for a period of 4 months, and measure overlap based on an extended network of 77 million users, to ensure minimal sampling errors.

We observe a strong relationship between local topology and tie-level behaviour, with some temporal features outperforming communication intensity in overlap prediction. Indeed, the number of bursty cascades, differences in daily behaviour and temporal stability play large roles in our models. We find that communication intensity is one of many characterizations of tie strength for which the Granovetter effect is observable.

Keywords: Complex Networks, Social networks, Tie Strengths, Granovetter effect, Temporal Patterns, Communication Networks, Human Behaviour

Preface

This is a thesis about the analysis of social networks through communication data. Reading through years of research, experimenting with data and analyzing anonymized call logs of millions of people, at some point it feels inevitable not to start noticing patterns in your own communication activity. Of course, the setting is substantially different from that which I research here: in recent years the available channels of communication have increased, and being in different continents than most of my family and friends implies not only different channels, but also time zones and activity patterns. Nevertheless, the fact is that I am still studying a phenomena in which I am deeply ingrained, and studying temporal features of human communication makes it impossible not to start paying attention to your own social network. Here, I want to thank all the people who have helped me through these years:

My parents, Laura and Javier, who have been so supportive and loving for so many years. And my sister, Laura Bertha, who has always been there to joke and support. This is truly an achievement of the three of them.

My supervisor, Mikko, and my good friends and colleagues from the Complex Systems group, whose advise and comments have been incredibly helpful.

My dear friends Sofía and Paola, whom I have known for so many years, and whose conversations have been incredibly valuable for getting perspective.

My wonderful flatmates and friends, Olli, Ulla and Frida, who have been so helpful and supportive making the quotidian feel joyous and fun.

Tonttu, the *good boy* who in a just a few months has mastered the ability of cheering me up - and just making life better-, via tail-wagging.

Raúl, who makes it all worth it. You are so incredibly special to me, and I am so happy with the life we are building together.

Helsinki, 27.05.2019

Javier Ureña Carrión

Contents

1 Introduction

Social networks seem to be everywhere nowadays. During the last few decades, the advent of digitization, the ubiquity of mobile phones and the prominence of large online platforms such as Facebook and Twitter in shaping the modern political landscape, from the Arab Spring [18] to Brexit [47], have all turned the term *social network* into a fixture of the current lexicon. For these, and many more reasons, it's easy to forget that the large-scale analysis of social networks is a development only a few decades old [48, 36, 17], while the concept of society as a network originated from sociological perspectives in the 20th century [15]. Certainly, social network analysis is a field in continuous renewal as a tool to understanding social systems [42, 17]. In this thesis we will explore social networks -not as online platforms, but as webs of relationships and connections-, through the analysis of mobile phone call records; that is, by understanding the system generated by person-to-person communication interactions.

Yet in social networks not all relations and interactions are the same, as people live wrapped in a series of highly diverse ties, whether family and friends, or colleagues and acquaintances from all walks of life. Indeed, these different *tie strengths* have been theorized to play a role in social networks since at least mid-20th century [15]. And in recent years, the use of data-based methodologies has allowed us to associate different tie strengths to differences in network-level behaviour [36, 23, 44]. But what makes a tie *strong*?

Social ties are varied and multifaceted, and from a quantitative perspective, particularly hard to measure [49]. Precisely because people value relationships differently, it is not a trivial concept to generalize to a network-level standard [49, 50]. There have been many approaches as to how to measure tie strength, from communication intensity-based measures - such as the total number of calls [36], to theoretical frameworks based on sociological theories [15, 35] and self-reported strength [49, 43]. In this thesis we take the sociological theory proposed by Mark Granovetter [15] as a basis to assume that *the strength of a tie is a latent variable associated to the community structures around the tie*, one that manifests in network topology. Based on this assumption, we may analyze which behavioural patterns are linked to the network's local topology allowing us to *identify which markers are*

more associated to tie strength via community structures.

This definition is not all-encompassing and absolute and we do not deny that strong ties may exist without community-structures around them. The use of this definition, however, does allow us to establish a general framework with a simple point of comparison for different temporal features. This way, we may explore diverse temporal aspects of human communication, while being able to determine whether they provide additional information that communication-intensity measures do not.

This thesis will thus be structured in a way that allows us to explore different aspects of human communication and social networks derived from communication data. This is, however, not a trivial task - our understanding of human communication patterns has rapidly evolved during the past few years [36, 17, 21, 29]. For this reason, we first include a general chapter on social and communication networks, followed by two chapters on distinct ways to understand how humans communicate in time: first, different ways of analyzing the linear sequence of events of human communication; second, how daily and weekly patterns also determine human interaction. In the final chapter we use all these temporal features to *predict* network topology given the temporal features we analyzed. In more detail:

Chapter 2 provides a theoretical background to the study of social networks, including the analysis of *communication networks*, which capture communication patterns of an underlying social network, and have become one of the primary tools for understanding social networks [17]. We discuss how to construct networks from call log data and the impact of different modelling choices on the resulting network (2.2.1). We then explain the theory of how *tie strengths* have been coupled to the topological *embeddedness* of a tie, and how the embeddedness can be measured via *edge overlap*. This is followed by an analysis of the effect of using CDR data on the network estimates of topological overlap. As we will see, using the data of a mobile-phone operator constitutes a form of sampling from a full communication network, and this sampling usually yields biased estimates of overlap. Since overlap is such a central measure of this thesis, we analyze from a sampling-theoretical perspective how sampling bias might affect overlap. Last, we cover a way to measure how overlap changes in time, and how this compares to measuring overlap from a static perspective.

The next Chapter 3 is devoted to a *linear* understanding of time: how different features might be derived from the sequences of calls between two people. Indeed, the study of the sequences of calls has been widely studied form different perspectives, some of which we explore and compare to network topology. First, we analyze the Inter-Event Time distribution, or the distribution of time between two calls. There, we'll see that many call sequences are *bursty*, or characterized by irregular patterns where many *bursts* of calls might be followed by long waiting times. In fact, we'll see that burstiness can be expressed in a wide array of manners, so we'll compare different expressions of burstiness to overlap. Last, we will focus on measures of temporal stability, of measures that attempt to capture *when* calls take place during the observation window.

Then, Chapter 4 focuses on how individuals act under a *cyclical* understanding of time, following daily and weekly patterns. We structure this chapter into two main parts. For the first part, we will compare the daily outgoing call distributions to analyze whether there is a correlation between the strength of a tie and the daily activity of people. Then, we inspect if there are certain times of the week at which people communicate that might be revealing of a community around them. To do so, we first examine communication patterns themselves, and then propose a clusterization procedure that allows us to find which groups of hours capture overlap variation.

For Chapter 5, we use the explored variables for the task of overlap prediction. As previously stated, we assume that both network topology and tie-level behaviour are expressions of tie strength, so we develop a series of experiments in order to find which variables show stronger links to network topology. We follow four scenarios: first, we use each variable separately for overlap prediction; then we use dual-predictors including every variable plus communication intensity in order to compare the roles of these two variables. Last, we create a full-variable model to predict both overlap and a measure of temporal overlaps.

2 Social and Communication Networks

The term *social networks* refers to graphs that attempt to capture the "notion that individuals are embedded in thick webs of social relations and interactions" [8], which at the same time is an effort to uncover how people, functioning as individuals, create relationships and community structures that scale to form societies [8]. The study of social networks originated in sociology during the last century [8, 29], and has evolved in recent years due to increased computational power and access to data. Certainly, the digital footprint generated by people in their daily lives has granted access to a vast amount of information that allows us to understand, model, and quantify the different patterns in which people behave and maintain social structures, as well as how such structures change over time [32, 17]. Social networks may be defined in a varied number of ways, as human relationships are varied themselves. For instance, we may define a social network based on kinship and relationship (such as friends, family and acquaintances), yet other types of human interactions might have other dynamics, such as collaboration networks in scientific research, or sexual relationship networks [44, 8].

The focus of this thesis will be on empirical communication networks, which are mostly derived from recorded human interactions via electronic devices such as phone calls, text messages or other types of instant messaging services. This type of network gained prominence during the last decade as both the accessibility and availability of mobile phones, computers, and social platforms increased [44]. Indeed, these networks have been crucial for the broader understating of social networks in general [44], as it seems that communication networks do capture many dynamics of the underlying social relationships [17]. We will analyze a particular type of communication network, obtained from Call Detail Records (CDRs), which are logs used by mobile phone operators to charge their customers.

In this chapter, we will discuss some aspects related to the construction of social networks from CDR data, as well as analyze common properties and characteristics normally observed in social networks. Now, since human communication occurs in time, the modeling of social networks might differ significantly if temporal factors are taken into account or not. For this reason, our discussion will mostly focus on different approaches that may deal with temporal aspects of data, yet most of

our results will rely on a *static* approach; that is, where time is disregarded and the network is assumed to behave similarly in time. We do, however, introduce an approach to modelling temporal topological changes in Section 2.2.4, but this is only for exploratory purposes. In the following chapters, however, we will analyze how temporal aspects of people's behavior are related to the broader static network.

2.1 Definitions

Networks are mathematically represented by *graphs*, defined as a pair of sets $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of N vertices or nodes, and $\mathcal{E}$ is a set of L links or edges of the form $e_{ij} = (i, j)$ joining the elements $i, j \in \mathcal{V}$. A graph is call *directed* if the order of the nodes in an edge is relevant, that is $e_{ij} \neq e_{ji}$ for $e_{ij}, e_{ji} \in \mathcal{E}$, or *undirected* otherwise. Edges may then be represented as an adjacency matrix A of size $N \times N$, where entry $A_{ij} \in \{0, 1\}$ is a binary variable representing either the presence or absence of an edge.

For social networks, it is common to define nodes as individuals, and edges as connections between individuals; however, the nature of the connections is not trivially defined. A *weighted* network, for instance, assigns values w_{ij} that represent the intensity of an edge: number of calls for communication networks or number of work collaborations for academic networks. Part of the objective of this thesis is to analyze how different weight definitions affect are related to network structure.

The *degree* k_i of node i in an undirected graph is equal to the number of incident edges and is a basic measure of edge connectivity. In the case of weighted networks, the *strength* s_i of a node is equal to the sum of the weights of the incident edges, so that $s_i = \sum_j w_{ij}$ [7]. In directed networks, both degree and strength are defined according to the direction of the edges, so that the *in-degree* is defined as the number of edges directed towards the node, and the *out-degree* is the number of edges directed from the node.

2.2 Empirical social networks

During the last century, the study of human interaction as networks evolved from theories developed from a sociological perspective into computational social sciences

[8]. The use of Call Detail Records (CDR) amplified the scope and insight into communication dynamics and social networks, particularly during the last fifteen years [36, 17, 29]. This type of digital record, which contains basic information of mobile phone interactions - such as identifiers for the caller and callee, duration, cost and type of interaction, among others- gained prominence as it enabled the analysis of social networks at regional, national and planetary scales [8] during time-frames spanning months and years [36, 29]. This type of data has therefore been used for a wide variety of fields, from social network structure [36] to human mobility patterns [14], prediction of tie formation [32] and spreading phenomena [25]. In addition, the use of this data has helped understand how social networks differ from random networks in terms of topology, connectivity, and temporal dynamics.

The analysis of social and communication networks has also evolved following different understandings of the temporality of the data. Earlier studies relied mostly on the analysis of *aggregated* networks, where the network would be assumed to be static in time, and link weights would be defined in terms of intensity, either by the number of interactions or total call length or other related measures. These studies revealed prominent characteristics of empirical social networks, such as community structures [36], fat-tailed degree distributions [12], and small-world phenomena [23]. Further studies shifted their focus to the temporal patterns of communication events, characterizing human activity as *bursty*, with periods of intense activity and longer periods of inactivity [13, 21]. As we will see in 3.2, burstiness plays an important role in information diffusion [23] while complicating the analysis of the persistence of ties [30]. Another line of studies has focused on daily patterns and circadian rhythms [4, 44], or analyses of how the time-allocation patterns of individuals [43, 33].

For now, we shall examine some of the characteristics that differentiate social from other types of network data, starting from ways to construct social networks from CDR data, followed by defining tie strengths in social networks, and seeing from a more theoretical perspective how sampling of networks affects statistical estimates.

2.2.1 From CDRs to networks

When creating a network from CDRs, different choices and construction rules have a significant impact on the resulting network. These choices involve, for instance,

the temporality of the data, the aggregation method, the required preprocessing and the effect of sampling nodes from a full network of people calling each other, which corresponds to dealing with data from specific operators that have a limited market share. In this section, we will briefly cover some of the methods that, to the best of our knowledge, are commonly used when dealing with CDRs, and some of the implications on the resulting network [44].

A common approach to creating networks is establishing an undirected weighted link between two nodes if there have been reciprocal calls, and defining weights as some form of aggregated communication intensity, such as the number of calls, the total call length or the average call duration. The reciprocity condition is used as a way to ensure familiarity between individuals and filter either spam or commercials calls. In this sense, it is related to a conceptual problem of inferring a latent social relationship given a communication event [11, 49]. The weights are commonly used as proxies for tie strength, depending on the application or purpose of the analysis [44]. Needless to say, both reciprocity and link weights are thus central themes to this thesis.

This approach, where a reciprocal call grants a link of weight defined by intensity constitutes a static approach to social networks and carries several premises inasmuch it neglects the temporality of the data. Most importantly, this method assumes that any event is just as likely to occur at any moment, implicitly understanding the activation times as a homogeneous Poisson process where there are no correlations between calls or temporal patterns [25, 29]. As we will see in the section 3.2, the burstiness of human communication is incompatible with a Poissonian model [32, 23]. A further assumption that stems from this static approach is that the underlying social network is static itself, while the reality is much different: neither nodes nor ties are persistent, as social ties are created and destroyed, people meet new people, students change social circles, and people fulfill their life cycles. Indeed, social ties exhibit not only highly varied lifespans of different intensities but also *memory*, where old links are more likely to persist than new ones [39]. In the end, communication networks involve at least two interwoven layers of temporality, first regarding the observed events, and second, the underlying dynamic relationships.

Different methodologies may be used to deal with either of these two temporal

aspects of communication networks, particularly depending on the time span of the source data. A first approach is to aggregate contacts over smaller periods of time, by splitting the data into consecutive time windows and obtaining a sequence of static networks, and study how these networks change [17, 29]. Time windows are, however, still static graphs, which can be a problem for certain types of data. In communication networks, the time between calls may be large (spanning several weeks or months, in some cases), resulting in temporal dependencies that are difficult to capture by this approach.

Other techniques use large longitudinal data, spanning years in some cases, to work around some of these issues. For instance, [32] uses a data set spanning 19 months and divides it into three distinct time windows, where the first and last serve as validation data that helps determine whether ties that have either died or been created in the central observation period. It is then possible to establish which ties from the first two periods persist in the third, and which ones which did not exist in the first period exist in the second and third. In other words, the use of an extremely large dataset enables us to analyze which ties are newly created (destroyed), and which characteristics and behaviors led to their creation (destruction).

In addition to temporality, the source and sampling of the CDRs are also relevant for constructing social networks. Because CDRs are obtained from telecommunication operators, it is common for the researcher to have only partial information corresponding to the market share of the operator, which constitutes a sample from a complete national (or, in a strict sense, global) communication network. For the purpose of this thesis, we will refer to the individuals who use the services of our operator who has provided data as *company users*, and *non-company users* to subscribers of operators who have not provided us with their data. This distinction is relevant since CDRs contain the complete history of company users including interactions with non-company users. In practice, it is common to filter non-company users, assuming that each node is independently sampled with probability equal to the market share [36]. This is done to use complete behavioral data of users, as we have no information of how non-company users behave among themselves, yet using only company users does guarantee observing full tie-level activity . As we will see in Section 2.2.3, this step carries non-trivial implications on the resulting network.

2.2.2 Tie Strengths and Topological Overlap

The adoption of call intensity as link weights has been mostly due to practical reasons, such as the involvement of temporal and economic commitment to ties [36], and has been interpreted as the strength of relationships when more data, such as surveys of emotional intensity, is unavailable [50, 43]. One of the main goals of this thesis is to understand other behavioral characterizations and features that illustrate the strength of a tie. Indeed, the same call intensity might occur in a myriad of situations. Take, for example, a tie with 40 calls. Such a scenario may occur in a period of three days or of two months, and in both cases that situation might happen during the weekend, at working hours, or be evenly spread during a period. In all these cases, then, assuming that all ties with 40 calls have the same strength might hinder information, instead of enlightening about the underlying relationship.

During the last century, sociologists developed an influential theory that associated tie strength to network topology, stating that strong ties between people are embedded in local community structures, while weak ties serve mostly as bridges between communities. In particular, Mark Granovetter's 1973 article "The Strength of Weak Ties" [15] became highly influential in the literature by proposing not only that strong ties are located in communities, but that information diffusion at a network level was mostly possible through weak ties by facilitating inter-community interactions. In social network theory, the embeddedness of a tie is usually measured via overlap, which can be regarded as a tie-level clustering coefficient. For $i, j \in V$, and $\mathcal{N}(i)$ the set of neighbors of node i, we define topological overlap as the Jaccard similarity between the sets of neighbors

$$O_{ij} = \frac{|\mathcal{N}(i) \cap \mathcal{N}(j)|}{|\mathcal{N}(i) \cup \mathcal{N}(j)|} \tag{1}$$

This expression takes values between zero and one, corresponding to the cases when there are no common neighbors, and when the sets of neighbors are equal, respectively. Figure 1 depicts a visual example of topological overlap in a network. We can define overlap in terms of the number of common neighbors n_{ij}, and the node degrees k_i and k_j [36]

$$O_{ij} = \frac{n_{ij}}{(k_i - 1) + (k_j - 1) - n_{ij}} \tag{2}$$

Thus defined, overlap has been used to observe the *Granovetter effect* or the increase of the strength of a tie with overlap. Indeed, [36] observed this effect in CDR data for link weight defined as the total number of calls. More recently, Miritello, et al [29] observed the same effect in a dynamical scenario, studying the evolution of overlap in time windows around the creation and destruction of ties. Their results suggest that the process of tie creation (or decay) between two people entails topological changes in time, where overlap increases even before a tie is created, and does so according to the strength of the tie (here again defined in terms of communication intensity). As Miritello's study suggests, topological changes occur dynamically, which points to the necessity of incorporating temporal aspects into the analysis of overlap. This poses more challenges, as long times between consecutive calls might be confused with both tie formation and decay, a difficulty that - to the best of our knowledge, is most commonly dealt with by analyzing longer periods of time.

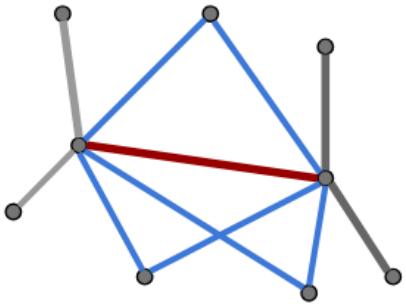


Figure 1: Visual example of topological overlap around a tie (red). We use a local community structure around a tie by analyzing the set of three common neighbors (blue ties) and the set of four non-common neighbors (grey ties). In this case, overlap is thus $O = \frac{3}{3+4} \approx .43$.

Despite the conceptual importance of tie strengths, their characterization is not universal and may be context-specific. From a sociological perspective, Granovetter defined it in his seminal work as a "possibly linear" combination of time, emotional intensity, intimacy and reciprocity [15], yet even under this scenario, emotional intensity and intimacy are not straightforward concepts to measure. While it is

rather common to define strength as communication intensity, this is not the case for all studies. Other possibilities include the use of survey-data as a basis for emotional closeness, as [43] did for ego networks, although this approach is unfeasible for large-scale CDR data, where surveys are prohibitively expensive. More recently, a study by [35] used Granovetter's multidimensional definition with the purpose of predicting tie decay. They created a set of variables that would serve as proxies for time, intensity, intimacy and reciprocity, thus bypassing a simple characterization of tie strength and linking specific variables to tie decay, assumed to occur in weak ties. Interestingly, they found that temporal features -such as the time spent without communication events, are the best predictors for tie decay.

In this thesis, we intend to explore different measures of tie strength that go beyond communication intensity by analyzing how these different measures correlate to, and are able to predict, topological overlap. In this sense, we will assume that tie strength is a latent variable that is positively correlated to a tie's embeddedness - in line with Granovetter's hypothesis, and that a better predictive power for overlap implies a higher tie strength.

There is a main drawback to this approach. As previously mentioned, the *strength* of a tie is an elusive, multidimensional and context-specific concept. Using embeddedness as a proxy for strength will, by itself, misrepresent ties that are "strong" by other measures. Indeed, defining strength based on communication intensity measures - time spent calling and mentions for Twitter networks, [38] found *strong* ties that are long-range; that is, that are otherwise separated by a path that is longer than one common neighbour. Our approach to tie strength thus does not intend to be all-encompasing, but to offer a network-inspired baseline that links temporal behaviour with topology. In this regard, then, ours could be thought of as a topological tie strength.

For now, we will continue our chapter on communication networks by exploring how the sampling scheme -data obtained from a single operator-, affects our measurements of overlap. Indeed, there are many non-trivial steps that affect network construction, and we use a mixture of a network of company-users for temporal behaviour, while a an extended network of company and non-company users for our topological estimates - a decision that we justify in the following section.

2.2.3 Effect of sampling on overlap

When dealing with CDRs, or any other type of massive social network data, it is common for our data to be a sample of the underlying communication network; for instance, if only one mobile phone operator provides its data, we will be dealing with a sample that is the size of the market share of this operator, even if people communicate with each other regardless of their service provider. All in all, CDRs are themselves a sample of human communication, being just a specific channel and not a full observation set of human interaction. Different sampling schemes and different estimators may result in wildly different results [27]. Since topological overlap is such a central concept in this thesis, we find it necessary to guarantee its correct measurement. As we will see, some common assumptions and preprocessing steps may have an unwanted effect in overlap estimates, a reason why we opted to measure overlap using a dataset with both company and non-company users. In this section we will discuss the effect of sampling from a theoretical perspective.

In network sampling schemes it is common to assume that nodes (or edges) are sampled independently with selection probability p [27, 36]; in other words, it is common to follow a Bernoulli sampling scheme. This might be a strong assumption in empirical networks, particularly for the analysis of triangle-structures (where there are three edges that connect three people), as it implies that the reasons why any two people might have the same operator are completely independent, ignoring cases where all or most of the family members are clients of the same company, where there are regional differences of market share, or company-level contracts with an operator, for example. In this section we shall analyze the effect of independent sampling into our estimates of overlap from a theoretical perspective, followed by an analysis of our data in section 2.3.1. As we will see, our study suggests that nodes in our data are not independent and randomly selected, but contain systematic bias that is related to the local topology of the network itself.

For the purpose of clarity, we'll first exemplify how sampling from networks is related to obtaining estimates. Consider how to estimate the population size under a random sampling scheme where nodes are selected independently with probability $p = 0.2$, the market share of the company in our data. Let N be the real size of the population, and let n be the size of the sampled population. In this case $\hat{N} = \frac{n}{p} = 5n$

is the Horvitz-Thompson [27] estimator for N: we assign each sampled node a weight associated to their inverse election probability, so that if we sample a person with probability p, this person accounts to $1/p$ of our estimates for the full population. Under the same assumptions, the probability of observing a link is p^2, and the probability of observing a triangle is p^3. Our estimators for the number of links and number of triangles would be, respectively, $\hat{L} = \frac{l}{p^2} = 25l$ and $\hat{T} = \frac{t}{p^3} = 125t$, for l the observed number of links and t the observed number of triangles. In this sense, the underlying call network would contain 25 more times of links than the observed network, and 125 times the observed number of triangles.

Let us now consider how to estimate overlap between two nodes. Under the Bernoulli sampling scheme, the estimator for overlap O_{ij} found in equation 2 is unbiased. To see this, we will divide the sets of neighbors of the two nodes into three mutually exclusive sets: one of common neighbors, and two of nodes that are only neighboring either i or j. Let $\mathcal{N}_{ij} = \mathcal{K}_i \cap \mathcal{K}_j$ be the set of common neighbors, $\tilde{\mathcal{K}}_i = \mathcal{K}_i \setminus (\mathcal{N}_{ij} \cup \{j\})$ the set that includes the neighbors of i and not j, and $\tilde{\mathcal{K}}_j = \mathcal{K}_j \setminus (\mathcal{N}_{ij} \cup \{i\})$ the neighbors of j and not i. By design, we have that $\mathcal{N}_{ij} \cap \tilde{\mathcal{K}}_i = \emptyset$, $\mathcal{N}_{ij} \cap \tilde{\mathcal{K}}_j = \emptyset$, and $\tilde{\mathcal{K}}_i \cap \tilde{\mathcal{K}}_j = \emptyset$. This allows us to write O_{ij} (equation 2) as

$$O_{ij} = \frac{|\mathcal{N}_{ij}|}{|\tilde{\mathcal{K}}_i| + |\tilde{\mathcal{K}}_i| + |\mathcal{N}_{ij}|} \tag{3}$$

Now, let us consider the commonly modelled sampling scheme of independent Bernoulli sampling for a set of nodes $\mathcal{S}$. If we do independent trials, each one with probability p, then the number of sampled nodes $S \sim \text{Binomial}(|\mathcal{S}|, p)$. Taking this into account, we define random variables that capture the number of observed neighbors, $N_{ij} \sim \text{Binomial}(|\mathcal{N}_{ij}|, p)$, $K_j \sim \text{Binomial}(|\tilde{\mathcal{K}}_j|, p)$ and $K_i \sim \text{Binomial}(|\tilde{\mathcal{K}}_j|, p)$. For the sets of non-common neighbors, we simplify our expression by adding the two variables, which are now distributed $K_{ij} = K_i + K_j \sim \text{Binomial}(|\tilde{\mathcal{K}}_i| + |\tilde{\mathcal{K}}_j|, p)$. Our estimator for overlap is then $\tilde{O}_{ij} = \frac{N_{ij}}{K_{ij}+N_{ij}}$, which does not have an easy closed-form distribution. We will then prove that it is an *approximately* unbiased estimator by using a Taylor expansion around its expected value.

Let $X \sim \text{Binomial}(x, p)$ and $Y \sim \text{Binomial}(y, p)$ be two independent variables. We have,

$$\mathbb{E}\left(\frac{X}{X+Y}\right) \approx \frac{\mathbb{E}(X)}{\mathbb{E}(X+Y)} - \frac{Cov(X, X+Y)}{\mathbb{E}^2(X+Y)} + \frac{\mathbb{E}(X)\mathbb{V}(X+Y)}{\mathbb{E}^3(X+Y)}$$

These expressions are equal to

$$\mathbb{E}(X) = xp$$
$$\mathbb{V}(X) = xp(1-p)$$
$$Cov(X, X+Y) = Cov(X, X) + Cov(X, Y) = \mathbb{V}(X)$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left(\frac{X}{X+Y}\right) &\approx \frac{xp}{xp+yp} - \frac{xp(1-p)}{(xp+yp)^2} + \frac{xp\left(xp(1-p)+yp(1-p)\right)}{(xp+yp)^3} \\
&= \frac{x}{x+y} - \frac{xp(1-p)}{p^2(x+y)^2} + \frac{xp^2(1-p)(x+y)}{p^3(x+y)^3} \\
&= \frac{x}{x+y} - \frac{x(1-p)}{p(x+y)^2} + \frac{x(1-p)}{p(x+y)^2} \\
&= \frac{x}{x+y}
\end{aligned}$$

It follows that $\mathbb{E}(\tilde{O}_{ij}) \approx \frac{|\mathcal{N}_{ij}|}{|\mathcal{K}_{ij}|+|\mathcal{N}_{ij}|} = O_{ij}$, which means that on average, the observed overlap is equal to the overlap of the complete communication network.

This result, however, still relies on the assumption that each node is sampled independently and with the same probability. We will now see that, if the selection probabilities of common neighbors are higher, this estimator for overlap is biased. Let us now assume that the selection probabilities satisfy $p_{N_{ij}} > p_{K_{ij}}$ for the case when $N_{ij} \sim \text{Binomial}(|\mathcal{N}_{ij}|, p_{N_{ij}})$ and $K_{ij} \sim \text{Binomial}(|\tilde{\mathcal{K}}_i| + |\tilde{\mathcal{K}}_j|, p_{K_{ij}})$.

Let $X_p \sim \text{Binomial}(x, p)$ and $Y_q \sim \text{Binomial}(y, q)$, and consider $0 < \alpha < 1$ such that $q = \alpha * p < p$. We may then follow similar steps as above to compute

$$\mathbb{E}\left(\frac{X_p}{X_p + Y_q}\right) \approx \frac{x}{x+y} + \frac{\alpha(1-\alpha)pxy}{(\alpha x + y)^3}$$

Since $\frac{\alpha(1-\alpha)pxy}{(\alpha x+y)^3} > 0$, it follows that different selection probabilities imply that the use of sampled data for overlap results in an overestimation for overlap. In the results section (2.3.1) we will use two different methods to see that, indeed, the selection

probability of common neighbors tends to be higher than for non-common neighbors, which results in biased estimates of overlap in the network of *company users*.

2.2.4 Temporal overlap

Overlap is a static measure of network topology, as it requires for us to use a time window ΔT where we can calculate overlap. Indeed, ΔT is commonly defined to be the whole observation window, as we previously did in this thesis. Nevertheless, recent studies have focused on the dynamic aspect of social networks: ties are created and die in spans of months, some individuals are *explorers* and some are *keepers* and thus renovate and change their social circles at vastly different rates [30]. As previously mentioned on section 2.2.2, Miritello, et al [29] have found through their *Dynamical Granovetter effect* that topological changes precede (succeed) tie decay (creation). When dealing with topological overlap from a static perspective, we do not take into account whether ties have decayed, and assume that if an event has been observed, then a tie exists. Indeed, since the goal of this thesis is to analyze how temporal behaviour is related to notwork topology, we believe it is important to investigate the effect of these temporal dynamics on overlap. In this brief section, we propose a method to analyze changes in overlap, exploring overlap not as a static number but as a time series, while also focusing on whether the sets of neighbors change during the observation window.

We approach this issue by defining a window ΔT that is smaller than our observation window. A common approach to analyzing temporal networks is to divide our observation period into subintervals, and calculate overlap (or any features of interest) during each period as if each subinterval constituted a static network [17]. This method, however, is prone to having abrupt changes, particularly as several ties (both for common neighbors and non-common neighbors) may not be present at different period of size ΔT.

Instead, we base our approach on the understating that there are *active* and *inactive* ties [29, 35], where active ties are ones where contact is likely to exist. Indeed, although an ideal situation would be to determine whether ties are active based on data itself, for smaller observation windows this is not possible [29]. We will thus determine the activity status of a tie by assigning a ΔT activation length to

each call. This allows us to measure overlap at different points in time defined by intervals of size ΔO^t, resulting in a time series of overlap values $\{O_{ij}^t\}_t$. We then define statistic on the time series. For instance, the mean temporal overlap $\hat{O}_{ij}^t$ is the mean overlap value of the overlap time series: $\hat{O}_{ij}^t = \frac{1}{n_t} \sum_t O_{ij}^t$, where n_t is the number of observations in the time series.

2.3 Results

We use CDRs obtained from a major operator in a European country with a market share of approximately 20% of the target population, obtained during the first four months of 2007 [23]. Our focus will be on ties for nodes that are company users, since we posses incomplete data regarding non-company users. However, since we know that using this network to estimate topological overlap results in biased statistics (results on section 2.3.1), we will compute overlap including non-company neighbors of our nodes, regardless of their provider. For the most part of this thesis, we will attempt to understand the relationship between a series of mostly temporal features and the topology of the static networks as a means to understanding which features could be used to predict tie strength.

We construct our static network by adding undirected links between nodes that have engaged in at least one communication event, and define link weight w_{ij} as the total number of calls during our observation period. We do not impose a reciprocity condition on our data because reciprocity is one the variables of interest, when our objective will be to analyze whether reciprocal calls are indicators of higher tie strength. We do, however, impose a limit on nodes that have a disproportionately large number of outgoing calls and no incoming calls, which we suspect correspond to customer services and not individuals.

Our company network has $\sim$ 6.5 million nodes and $\sim$ 26.4 million links. The extended network, which includes non-company users, has $\sim$ 75 million users. We structure our first results in the following manner: first, we justify our choice of using non-company users for measuring overlap by showing that in the network of only company users, we are more likely to observe common neighbors than non-common neighbors ($p_{N_{ij}} > p_{K_{ij}}$). Afterwards, we focus on static measures of tie strength, such as communication intensity and reciprocity; we finalize this results sections by

exploring how overlap varies in time.

2.3.1 Sampling and overlap

As seen in section 2.2.3, if there are differences in selection probabilities, then our estimates for overlap are biased. Here we will use two different approaches to show that the probability of sampling a common neighbor, $p_{N_{ij}}$, tends to be larger than the probability of sampling a non-common neighbor, $p_{K_{ij}}$. The first approach will be based on simulations, while the second will use a Bayesian framework to estimate $p_{N_{ij}}$ and $p_{K_{ij}}$. Our dual methodological choice reflects the fact that the set of common neighbors $|\mathcal{N}_{ij}|$ tends to be much smaller than the set of non-common neighbors, $|\mathcal{K}_{ij}|$, meaning that standard statistical tests for difference in proportions are unfeasible. For instance, for many links the observed number of common neighbors n_{ij} is rather small, invalidating tests that rely on the Central Limit Theorem, such as Pearson's χ^2 test. On the other hand, using exact tests such as Fisher's F might also be computationally expensive, as the number of non-common neighbors tends to be prohibitively large. Both simulations and Bayesian estimates attempt to correct for the differences in size of $|\mathcal{N}_{ij}|$ and $|\mathcal{K}_{ij}|$: the former by telling us how we expect our estimates of $p_{N_{ij}}$ and $p_{K_{ij}}$ to look like if they are equal, despite being sampled from unequal sets; while the latter framework by incorporating *prior* probabilities that smooth our estimates.

For the first case, we will simulate sampling a network from an extended network that contains both company and non-company users, where both common neighbors and non-common neighbors are sampled with the same probability. We will then obtain *sampled* estimates of $p^S_{N_{ij}}$ and $p^S_{K_{ij}}$, knowing that they are generated by the same number. We will then compare the distributions of our estimates for sampled and observed probabilities, and see whether there are differences. In more detail, given a link (i, j),

- n^c_{ij} is the observed number of common neighbors in the complete network that includes non-company users. That is, this number is equal to $|\mathcal{N}_{ij}|$.
- n_{ij} is the observed number of common neighbors in the company network. In other words, this is a sample obtained from N_{ij}

- k^c_{ij} is the observed number of non-common neighbors in the complete network, equal to $|\mathcal{K}_{ij}|$.

- k_{ij} is the observed number of non-common neighbors in the company network, a sample obtained from K_{ij}.

First we focus on the observed network, where we obtain estimates of selection probabilities for both common neighbors and non-common neighbors, via $p_{N_{ij}} = \frac{n_{ij}}{n^c_{ij}}$ and $p_{K_{ij}} = \frac{k_{ij}}{k^c_{ij}}$. Figure 2 (*left*) depicts the distributions of $p_{N_{ij}}$ and $p_{K_{ij}}$, which are visibly dissimilar. This is partly an effect of the small size of the set of common neighbors $|\mathcal{N}_{ij}|$: since it tends to be comparatively smaller than $|\mathcal{K}_{ij}|$ the estimates of $p_{N_{ij}}$ values are highly irregular. We know this to be an effect of the smaller size of $|\mathcal{N}_{ij}|$ as we can replicate it by sampling with known parameters. For each link we obtain the probability of selection of any (common and non-common) neighbor $p_{NK_{ij}} = \frac{n_{ij}+k_{ij}}{n^c_{ij}+k^c_{ij}}$, and sample both common neighbors and non-common neighbors with the same probability $p_{NK_{ij}}$. We use $p_{NK_{ij}}$ and not other values, such as $p = 0.20$, since there may be other factors that play a role in the selection probability, such as different market shares per geographical location or a more loosely-defined community structure that determines company affiliation. We used the sampled common and non-common neighbors to obtain the simulated selection probabilities $p^S_{N_{ij}}$ and $p^S_{K_{ij}}$, depicted on Figure 2 (*right*). This method illustrates that, even using the same selection probability, the comparatively smaller size of $|\mathcal{N}_{ij}|$ results in what seems to be higher estimates of $p_{N_{ij}}$, along with a relatively large number of zeros.

This still masks certain differences in distribution. We performed a Kolmogorov-Smirnov test to discern whether the distributions of the empirical and simulated values are different. This method tests for the largest difference in the cumulative distributions, and its test statistic D takes values in the interval [0, 1], where smaller numbers imply larger similarity in distribution. Our results show that the difference between the distribution of $p_{K_{ij}}$ and p^S_K is $D_K = 0.027$, while $p_{N_{ij}}$ and p^S_N show a larger difference, $D_N = 0.116$, both with p-values significantly smaller than 0.01. Since $p_{N_{ij}}$ and $p_{K_{ij}}$ cannot be replicated with the same mechanism, this suggests that they are, in fact, dissimilar. Figure 3 compares the empirical and simulated cumulative distributions for both cases, where the differences in the distributions

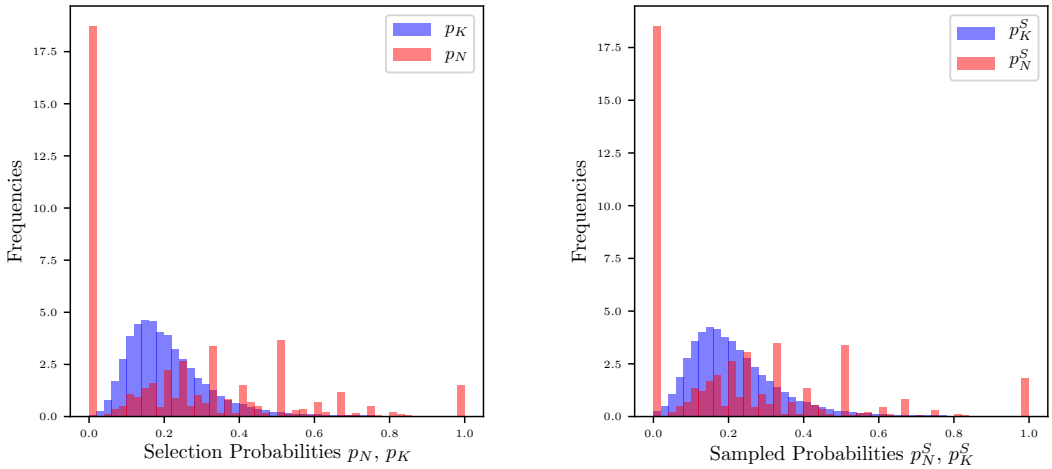


Figure 2: Empirical and simulated distributions of selection probabilities of common neighbors ($p_{N_{ij}}$, in red) and non-common neighbors ($p_{K_{ij}}$, in blue). (*left*) Empirical distribution, (*right*) Simulated distribution of selection probabilities, where both $p^S_{K_{ij}}$ and $p^S_{N_{ij}}$ were obtained with the same selection probability $p_{NK_{ij}}$. The simulation seems to capture the small-size effect of estimating $p_{N_{ij}}$, resulting in a distribution with a large number of zeros and large right-hand dispersion, while keeping $p_{NK_{ij}}$ constant for each $p_{N_{ij}}$ and $p_{K_{ij}}$.

of common neighbors are more evident. This graph also contains the first hint that $p_{N_{ij}} > p_{K_{ij}}$, since $p_{N_{ij}}$ consistently takes higher values than p^S_N.

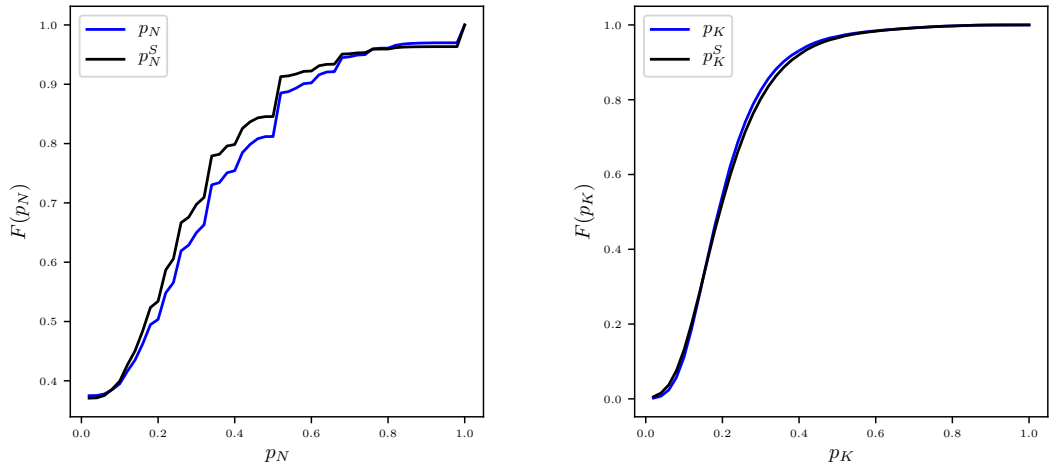


Figure 3: Comparison between empirical and simulated cumulative distributions of $p_{N_{ij}}$ (*left*) and $p_{K_{ij}}$ (*right*). While the sampling mechanism seems to yield a suitable distribution for $p_{K_{ij}}$, the simulated values for common neighbors are consistently smaller than the empirical $p_{N_{ij}}$.

For our Bayesian approach, we will test each specific link to see whether $p_{N_{ij}} > p_{K_{ij}}$. We model N_{ij} and K_{ij} as Binomial random variables, and choose Beta(α, β) as prior distributions for both $p_{N_{ij}}$ and $p_{K_{ij}}$. This choice of prior allows for us to compute an easy analytic posterior distribution. Let $D = (n_{ij}, n_{ij}^c, k_{ij}, k_{ij}^c)$ be the observed data, our posterior distributions have an analytic form

$$\begin{aligned} p_{N_{ij}}|D &\sim \text{Beta}(\alpha + n_{ij}, \beta + n_{ij}^c - n_{ij}) \\ p_{K_{ij}}|D &\sim \text{Beta}(\alpha + k_{ij}, \beta + k_{ij}^c - k_{ij}) \end{aligned} \tag{4}$$

We may interpret α and β, the prior hyperparameters, in terms of their smoothing effect on the posterior distribution. Indeed, $B \sim \text{Beta}(\alpha, \beta)$ has mean value $E[B] = \frac{\alpha}{\alpha+\beta}$, and thus our expected posterior estimates are $E\left[p_{N_{ij}}|D\right] = \frac{\alpha+n_{ij}}{\alpha+\beta+n_{ij}^c}$ and $E\left[p_{K_{ij}}|D\right] = \frac{\alpha+k_{ij}}{\alpha+\beta+k_{ij}^c}$. In other words, α and β have the equivalent effect of observing α "successes" of a binomial distribution, out of a possible $\alpha + \beta$ trials, and thus smooth our posterior probabilities. This is particularly useful for estimating $p_{N_{ij}}$, which we know tends to suffer from small-size effects.

We choose four sets of (α, β) pairs: (a) $(\alpha_1, \beta_1) = (1, 1)$, which corresponds to a uniform prior distribution. As we know, the market share of the company is around 20%, and this scenario assigns prior probabilities with expected values $E\left[p_{N_{ij}}\right] = E\left[p_{K_{ij}}\right] = 0.5$ which might be an unrealistic assumption. In particular, since in most cases the number of common neighbors tends to be much smaller than the number of non-common neighbors ($n_{ij} < k_{ij}$), the effect of this prior will be heavier on our estimate of $p_{N_{ij}}$. (b) $(\alpha_2, \beta_2) = (1, 4)$, a prior distribution with expected value $E\left[p_{N_{ij}}\right] = E\left[p_{K_{ij}}\right] = \frac{1}{5} = 0.2$, in line with our prior knowledge of market share. (c) $(\alpha_3, \beta_3) = (1, 6)$, a distribution with prior probability equal to 0.143, which will mostly penalize $p_{N_{ij}}$ and serve for comparison with other priors.(d) $(\alpha_4, \beta_4) = (0, 0)$, which does not correspond to a real prior distribution, but whose posterior exists when $n_{ij} > 0$ and $k_{ij} > 0$.

As depicted on Figure 4, the use of different priors has a smoothing effect on $E[p_{N_{ij}}|D]$, the posterior mean estimates for common neighbors, which seems to be rather sensitive to the choice of prior. Using a prior with mean value 0.2 implies that the distributions of $p_{N_{ij}}$ and $p_{K_{ij}}$ are now visually similar, even if the distribution of

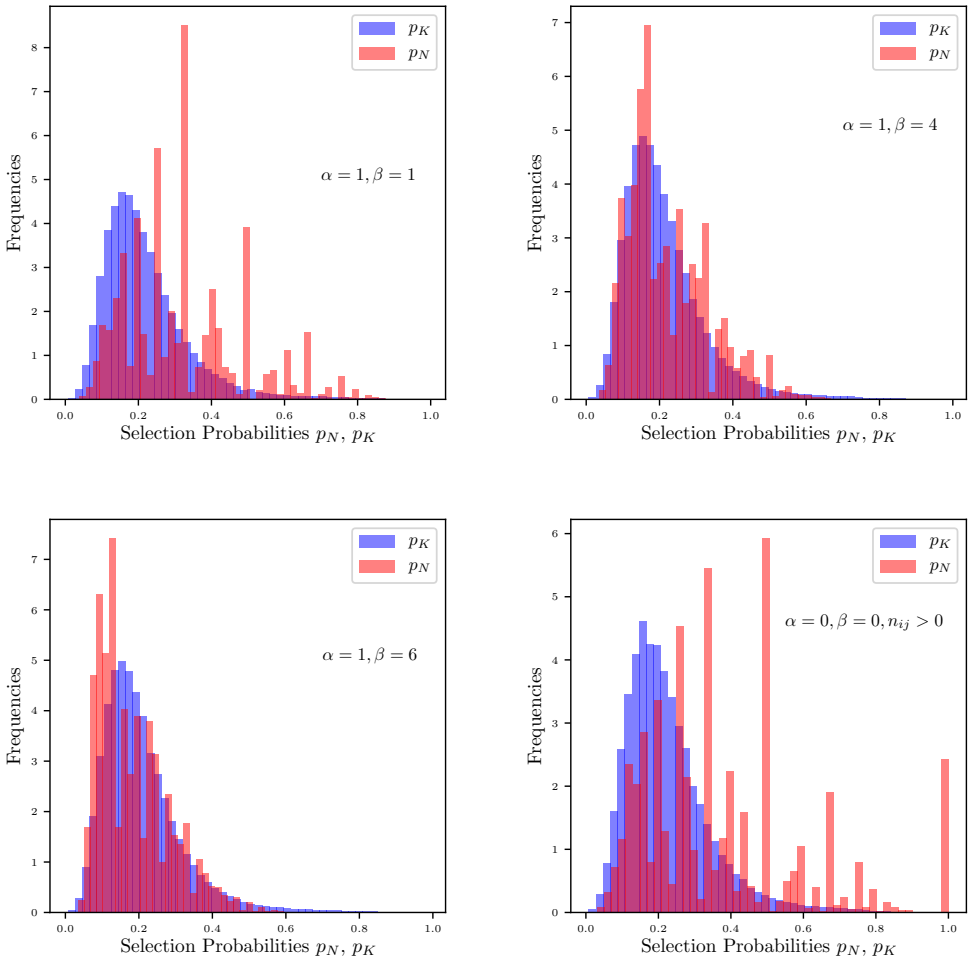


Figure 4: Posterior means for $p_{N_{ij}}$ and $p_{K_{ij}}$ using different priors.

$p_{N_{ij}}$ takes slightly larger values. Using a prior that penalizes small effects (with a mean value smaller than 0.2) confirms the sensitivity of $p_{N_{ij}}$, as it pulls the whole distribution towards smaller values than $p_{K_{ij}}$, even though we know that this prior is quite restrictive. On the other hand, the distribution for non-common neighbors, $p_{K_{ij}}$, seems almost unchanged for all three proper priors. As for the last plot, the condition $n_{ij} > 0$ introduces an important question: how do the selection probabilities compare *once we have observed common neighbors in the company network*. This question is of utter relevance, as it attempts to understand whether we have observed a common neighbor because we were more likely to observe it; that is, because of structural bias.

We now obtain $P_{SN_{ij}} = P(p_{N_{ij}} > p_{K_{ij}}|D)$, the probability that the selection of a

common neighbor is larger than that of a non-common neighbor, which we do by comparing the posterior distributions for each of our priors. It is not possible to generalize that *every link*'s common neighbors are selected with a higher probability, although we can confirm that this is the situation for a majority of links. Table 1 contains the percentage of links with probability $P_{SN_{ij}} > \alpha_c$ for different values of α_c and different priors (α, β). Likewise, Figure 5 shows the distribution of $P_{SN_{ij}}$ for different priors, where values greater than 0.5 serve as evidence for $p_{N_{ij}} > p_{K_{ij}}$. Although the choice of prior does have an effect on the proportion links where this holds, we notice that conditional on $n_{ij} > 0$, our results reveal a consistent bias for a higher $p_{N_{ij}ij}$. The choice of non-zero priors allows us to estimate posterior probabilities even when we have not observed a common neighbor, which might indeed be desirable in some cases, yet focusing on the case $n_{ij} > 0$ allows us to estimate the effect of systematic sampling bias. We therefore conclude that in the company network *once we have observed a link, we observe common neighbors with a higher probability than non-common neighbors.*

	All data			$n_{ij} > 0$			
α_c	(1, 1)	(1, 4)	(1,6)	(1,1)	(1,4)	(1,6)	(0,0)
0.5	0.69	0.47	0.36	0.82	0.65	0.55	0.66
0.55	0.64	0.41	0.32	0.78	0.59	0.49	0.61
0.6	0.58	0.36	0.28	0.74	0.53	0.44	0.56
0.65	0.52	0.32	0.24	0.68	0.47	0.38	0.51
0.7	0.46	0.27	0.2	0.62	0.41	0.32	0.45
0.75	0.4	0.23	0.16	0.56	0.35	0.26	0.39

Table 1: Percentage of links with probability $P_{SN_{ij}} > \alpha_c$ for different (α, β) priors. For all cases where $n_{ij} > 0$, more than 50% of links have a probability of observation grater than α_c.

These results justify our choice to use the extended network for estimating overlap, which we assume to be a full set of neighbors for our data. Indeed, not only does this avoid overestimating overlap due to structurally biased sampling, but it also reduces the number of zero-valued common neighbors.

2.3.2 Static measures of tie strength

Let us now focus on different measures of tie strength for a static network. First, consider the total number of calls w_{ij}, one of the most common measures of commu-

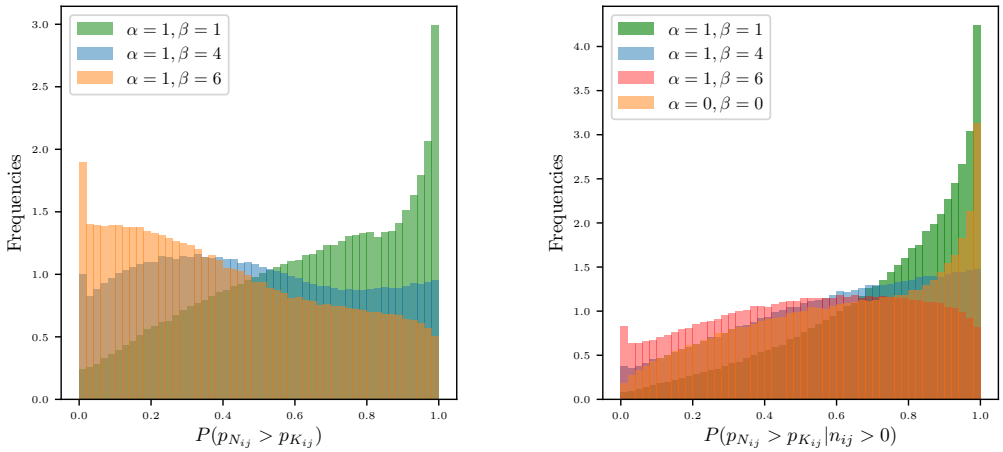


Figure 5: Posterior probability $P(p_{N_{ij}} > p_{K_{ij}}|D)$ using different priors, (*right*) for all links and (*left*) for links where we have observed a common neighbors in the sampled network, $n_{ij} > 0$. Focusing on cases where $n_{ij} > 0$ allows us to see that, on average, the selection probability of $p_{N_{ij}}$ is larger than $p_{K_{ij}}$, at least for the cases where we have already observed a common neighbor.

nication intensity [17, 36, 29]. As we have mentioned, this variable has been used to observe the Granovetter effect which we replicate here on Figure 6 using estimates for overlap in both the sampled network of company users, and the full network with non-company users. This shows that the average overlap tends to increase with the call intensity w_{ij}; in fact, the linear correlation is Pearson$(w_{ij}, O_{ij}) = 0.205$ and the rank correlation is Spearman$(w_{ij}, O_{ij}) = 0.41$. We notice, however, that there is a sharp decrease after w_{ij} reaches certain values, around 150 calls for the extended network, and 130 calls for the network of company users. This effect reflects that the distribution of w_{ij} is heavy-tailed. In fact, when considering and ordering of call intensity values, we get that this effect becomes almost negligible, which is also evident in the difference between the two correlation measures. As for what prompts overlap to fall so dramatically for larger weights, it has been shown that these links tend to spend most of their time speaking to each other [36]. This figure not only allows us to see the Granovetter effect, but also the effect of the sampling, as the company network consistently has higher overlap values.

We will now consider reciprocity, which Granovetter determined to be one of the main characteristics of tie strength [15]. Under a static approach, we define reciprocity as the ratio between outgoing and incoming calls for any node. Let $\vec{w}_{ij}$ be

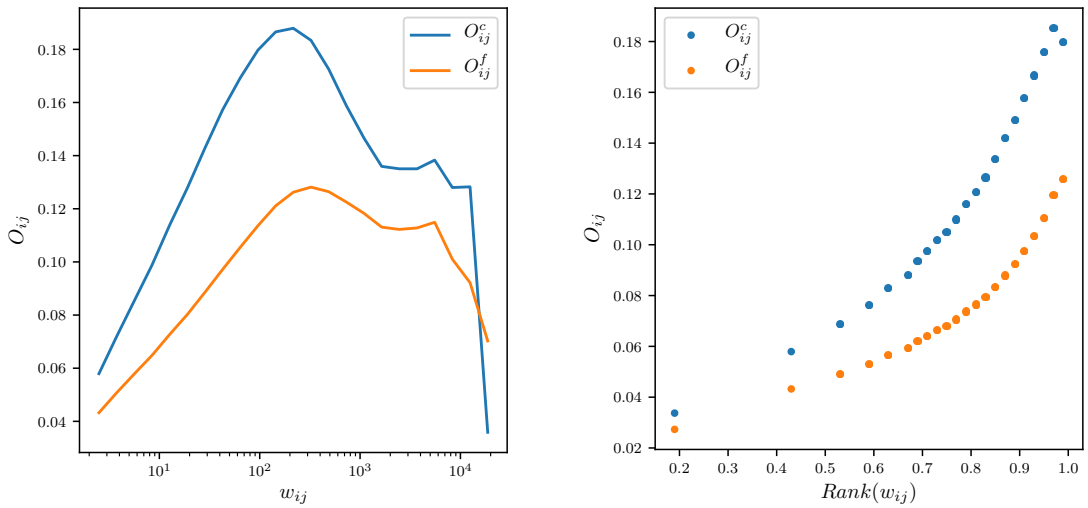


Figure 6: Visualization of Granovetter effect for call intensity (w_{ij}) defined as the total number of calls for overlap estimated with the sampled company network (blue) and with the full network including non company users (orange). (*right*) Average overlap conditional on link weight $\langle O_{ij}|w_{ij}\rangle$. The Granovetter effect, or the increasing trend of overlap in terms of w_{ij}, exists for a range of values, decreasing afterwards after a certain number of calls is reached. (*left*) Average overlap conditional on link weight rank, $\langle O_{ij}|Rank(w_{ij})\rangle$; since only small number of ties have extremely large call intensity, using rank allows to visualize how the Granovetter effect does indeed hold for most of the links.

the number of calls from i to j. We define the reciprocity between two nodes as [35]

$$r_{ij} = \left| \frac{\vec{w}_{ij}}{w_{ij}} - \frac{1}{2} \right|$$

Thus, if both users have placed the same number of calls to each other, we will have $r_{ij} = 0$, whereas if only one user has placed calls, we will have $r_{ij} = \frac{1}{2}$. In addition, because $\vec{w}_{ij} + \vec{w}_{ji} = w_{ij}$, our measure is undirected ($r_{ij} = r_{ji}$). Notice that under this definition our variable only measures reciprocity in terms of intent of call, meaning that a relationship may be reciprocal even if people don't call each other at similar rates. Indeed, if people have spent many hours talking with each other, that could be another proxy that the relationship is reciprocal, despite the fact that the calls are not. Figure 7 depicts the relationship between r_{ij} and O_{ij}, as well the relationship with w_{ij}, which does not seem to be large: overlap does not seem to change in any consistent manner with reciprocity.

This is slightly misleading, as computing the correlation between variables does yield a significant result: $\text{Pearson}(r_{ij}, O_{ij}) = -0.338$ and $\text{Spearman}(r_{ij}, O_{ij}) =$

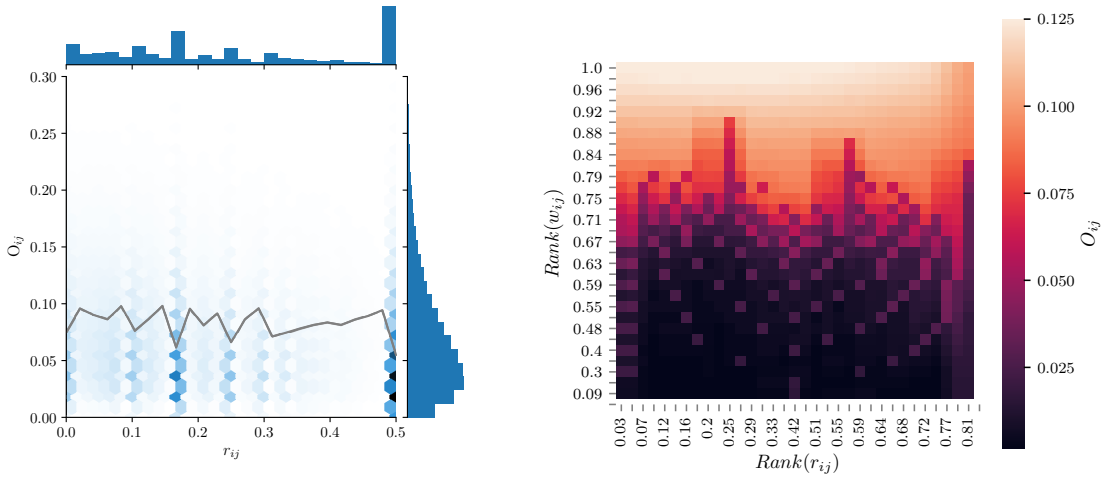


Figure 7: Relationship between call reciprocity r_{ij} and overlap O_{ij}, including (*left*) the joint empirical distribution with along with the average overlap for a fixed reciprocity value, $\langle O_{ij}|r_{ij}\rangle$ and (*right*) average overlap in terms of both r_{ij} and w_{ij}. Surprisingly, this variable does not seem to have large effect on overlap.

-0.412. This counter intuitive result stems from the fact that ties with no reciprocal calls ($r_{ij} = 0.5$) do tend to have a smaller topological overlap. In fact, conditioning on $r_{ij} < 0.5$ yields correlation coefficients that are close to zero: Pearson$(r_{ij}, O_{ij}|r_{ij} < 0.5) = 0.051$ and Spearman$(r_{ij}, O_{ij}|r_{ij} < 0.5) = 0.068$. On Figure 8 we compare the distributions of overlap and link weights depending on whether there have been reciprocal calls or not. While it is common to impose a reciprocity condition to the analysis of CDR data (that is, only analyze links where $r_{ij} < 0.5$), our results imply that links with no reciprocity may display not only positive overlap, but also the Granovetter effect for both call intensity and binary reciprocity. For this reason, we conclude that it might be more informative to adopt reciprocity as a variable instead of a necessary condition in a preprocessing step.

2.3.3 Temporal overlap

For our analysis of temporal overlap, the selection of both ΔT and ΔO^t will affect the behaviour and the granularity of the resulting time series. Here we will not perform an extensive analysis of the sensitivity of temporal overlap to the $(\Delta T, \Delta O^t)$ parameters, or of how we may interpret the resulting time series of temporal overlap

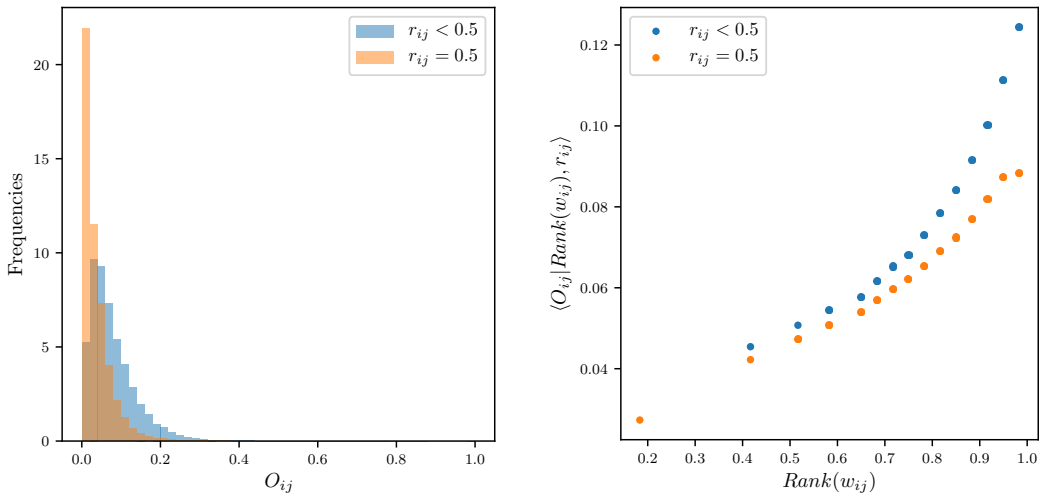


Figure 8: Relationship between binary reciprocity (for $r_{ij} < 0.5$ and $r_{ij} = 0.5$) and overlap. (*left*) Empirical distribution of overlap based on binary reciprocity, and (*right*) Granovetter effect for link weights conditioning on binary reciprocity. Although at a smaller degree, the links with no reciprocal calls also exhibit Granovetter effect.

$\{O^t_{ij}\}_t$. Instead, we will showcase a specific example of our data's behaviour for a fixed set of parameters: $\Delta T = 1$ month, and $\Delta O^t = 1$ week. Indeed, preliminary tests suggest that this choice of parameters exhibits some of the main differences between a static and temporal approach to overlap: a lower overlap distribution with a greater number of zero-valued links, high variation on the set of common neighbors, as well as a strong relationship to static overlap.

We compare static overlap and mean temporal overlap on Figure 9. It seems that both $\hat{O}^t_{ij}$ and O_{ij} have similar distributions - except of course for the fraction of zero-valued entries for $\hat{O}^t_{ij}$. Indeed, it seems that although both variables are tightly coupled, the temporal version tends to have slightly lower overlap values, which is expected, as we decompose interactions of a fixed set of common neighbors into different time periods. The question is now, *who* are the sets of common neighbors.

For this approach, we divide the sets of common neighbors in static overlap into three disjoint sets: neighbors that are connected to both nodes at all times, neighbors that are connected to both nodes at the same time for at least one observation, but not all; and nodes that are not connected to both nodes at any point in time (but are so in the aggregate). We find that only 1.32% of common neighbors are connected to both nodes all times; while 50.66% of neighbors are connected at some (but not

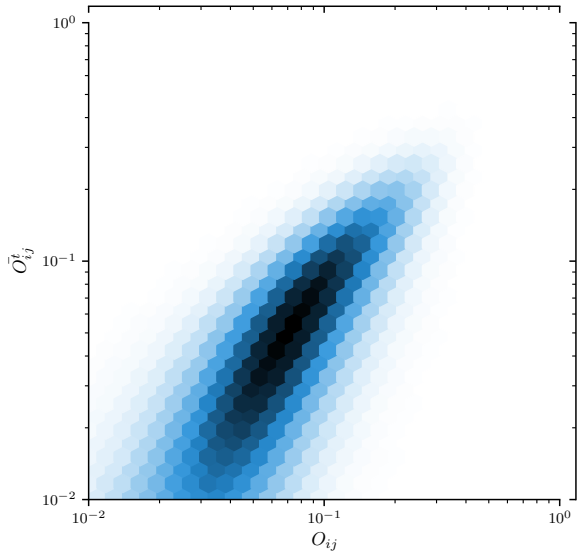


Figure 9: Joint distribution of topological overlap (O_{ij}) and mean temporal overlap $\hat{O}^t_{ij}$.

all) times, and 48.02% of neighbors are never adjacent to both nodes simultaneously.

The small set of common neighbors at all times is partially driven by the fact that only 7% of links have at least one common neighbor that is stable. We do find that these links tend to be high-intensity links, with an average number of calls of $\langle w_{ij}|\text{stable common neighbor}\rangle = 46.4$, as opposed to $\langle w_{ij}\rangle = 18.61$, which is the mean average number of calls for the whole population [1]. In addition, for these links we find that both overlap and mean temporal overlap are higher at 0.091 and 0.097, respectively, compared to the general population vales of 0.0528 and 0.033 for overlap and mean temporal overlap.

[1]Note, however, that this is an extremely heavy-tailed distribution, where 50% of the links have 3 calls or less.

3 Measures based on event sequences

While intensity or volume of communication is informative of tie strength (at least inasmuch the Granovetter effect is concerned), it alone does not uncover the myriad of ways in which such intensity takes place. Consider, for instance, the question of how often do people in a "strong" tie call each other? How regular are these calls? Are people more likely to call each other at a certain point of the day? If people have not communicated in a long time, has their relationship decayed? All of these questions are temporal in nature, and reflect some of the ways in which communication is time-dependent. We may, however, interpret time in different manners according to what we want to answer. In this chapter, we will understand time as sequence of events; that is, in a linear way. In the next chapter this will change, and we will approach time in a cyclical manner, where human activity is bound to 24-hour days and social schedules that determine working weekdays and leisure weekends.

So how does this "linear" approach to time reflect on the strength of a tie? We will divide this chapter into three major sections according to their theoretical foundation. The first one is based on the distribution of inter-event times (IETs); that is, how the waiting time between any pair of consecutive events is distributed. Next we will discuss burstiness. There is some conceptual overlap between this and the previous section as certain characterizations of burstiness are derived from the IET distribution; however, we discuss it in a separate section since not all approaches to burstiness are derived in this manner. Last, we will discuss variables related to *temporal stability*, in other words, variables meant for capturing any changes in activity or trends during our observation window.

As a relevant note, we would like to remark that communication patters -and therefore the IET distribution and any "linear" approach to time-, are at least partially determined by people's daily or weekly routine and habits [44]. The question of how the sequence of events and burstiness are related to people's routines has not escaped researchers, and has been explored in some scenarios [20]. For the purpose of this thesis, however, we shall distinguish this as two distinct approaches, at least in terms of how they are related to the Granovetter effect.

3.1 Modelling the inter-event time distribution

As previously mentioned, a key problem of using aggregated static networks is the non-trivial assumption that human interaction happens in a completely random manner, and can therefore be modelled as a homogeneous Poisson process [29]. If this were the case, it would be possible to model the number of events in terms of the size of the observation window T, and the intensity of interactions ρ, so that the total number of calls is distributed $Poisson(\rho T)$, and the time between any two events is exponential. Despite having these useful properties, evidence shows that this model does not hold for empirical networks. For instance, [6] estimated the inter-event time (τ) distribution to have a power-law behaviour $P(\tau) \propto \tau^{-\gamma}$ where $\gamma \approx 1$, thus rendering a heavy-tailed distribution. Here we discuss two distinct approaches to dealing with this distribution, as well as using different statistics to estimate its moments.

In more general terms, some authors have also determined that the inter-call time distribution between two nodes i and j may be expressed as as the product of the inverse average inter-event time, and a scaled universal function, so that $P(\tau_{ij}) = \frac{1}{\hat{\tau_{ij}}}\mathcal{F}(\frac{\tau_{ij}}{\hat{\tau_{ij}}})$ [13, 21]. This result was suggested by [9], who remarked that people distribute the time spent calling each other in a highly-skewed manner, and therefore proposed scaling inter-call times by a group-level average $\hat{\tau_G}$, where each group G is determined by the number of calls in the observation period. By using this method, the tails of the scaled distributions $P(\tau/\hat{\tau_G})$ collapse to a single heavy-tailed distribution; whereas the smaller scales of the distribution vary according to the used method.

For the purpose of this thesis, however, we wish to determine whether features of the IET distribution between two people are related to the network topology regardless of tie intensity measures (in the previous case, the number of calls which was used to make groups). We will now focus on how to estimate the moments of tie-level inter-call time distributions, particularly in the context of limited observation windows and the bias this induces, later on to use features of this distribution to find associations to overlap.

3.1.1 Estimating moments of the inter-event time distribution

The problem of estimating the IET distribution is, at its core, related to the problem of understanding tie creation and decay. We will contextualize this problem as follows: in the underlying society that generates communication networks, people maintain varied social circles, acquaintances and communities, and express part of these relationships in communication networks. These relationships, however, have distinct lifespans, and even their intensity and emotional attachment are not necessarily constant in time. Social networks are not static objects, and even if they were, we do not observe them directly, but via proxy variables: when observing CDR data, we are mainly observing communication events, not social relationships. It is therefore not trivial to infer that relationship from recorded interactions [29, 49]. One last problem that surfaces when using CDR data is that the timescale of the data may differ significantly from the timescales of social relationships, particularly for social relationships that remain inactive for long periods.

Given an observation window, the fact that we only observe data within that timeframe has several implications for network analysis, and the extent of any study performed on that data must be directly limited by this. It has been suggested that the use of longer longitudinal data might diminish the effect of observation windows. For instance, both [30] and [35] use longitudinal CDR data of 19 months in order to detect which ties have been formed in a period spanning 6 months, and which ties have decayed. All in all, it is still unfeasible to determine whether two people have just become acquainted or reactivated their tie after a long resting time. Conversely, it is not trivial to know whether a relationship has been severed or entered a period with no communication. However, the main goal is to study how peoples' social networks vary in time, which is possible to a fuller extent. Given that our study has four months' worth of data, we will focus on dealing with the constraints imposed by this observation window, and not on the underlying phenomena of tie creation and decay.

We will cover two different estimators for the moments of the IETs, in different attempts to deal with the observation windows. The first, which we refer to as the empirical estimator $\bar{\tau}^e$, is based on the empirical inter-event times, or the time between two consecutive events regardless of the observation window.

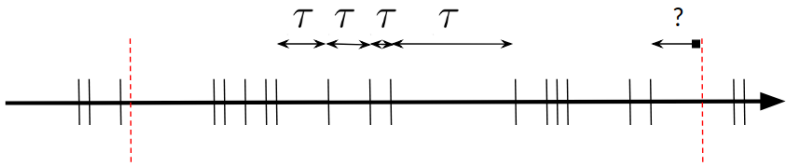


Figure 10: Graphic illustration of the effect of a finite-size observation window (in red) on the estimates of the IET-distribution. While it is possible to fully observe a set of inter-call times (τ), the observation boundaries imply incomplete observations. This is particularly difficult for low-intensity or extremely bursty behaviour, as we might not fully observe heavy-tailed scenarios.

Let $\{t_1, \ldots, t_n\}$ be the times for n observed calls between two nodes. We obtain the inter-event times as $\tau_i = t_{i+1} - t_i$, and obtain the moments according to this empirical distribution; for instance, $\bar{\tau}^e = \sum_1^{n-1} \tau_i$. As we can see on Figure 10, this estimator disregards some data on the observation window, and results in a linear cutoff at the end of the distribution [26]. In that sense, the empirical estimator results in a lower-bound of the real number.

The Kaplan-Meier (KM) estimator is a non-parametric method that attempts to correct for finite-size bias in survival functions. In rough terms, it takes into account *censored* data -intervals that have been observed up to a censoring time, and are thus assumed to be at least as long as its censoring time [26]. For the application at hand, we consider the inter-event time from the start (end) of the observation period and the first (last) event as a censored observation, allowing us to estimate the cumulative function of the inter-event times $P^{est}(\tau)$ with this additional information.

In order to estimate the mth moment of the distribution we use [26]

$$\mu_m^{KM} = \int_0^{\tau_{max}} \tau^m p^{est}(\tau)\delta\tau + \tau_{max}^m P_{\geq}^{est}(\tau_{max}) \tag{5}$$

Now, the use of the KM estimator implies certain assumptions from the data; in particular, that the generating process is a stationary renewal process. We know that the underlying social network is not static, and that some ties likely be created or destroyed in the period under study; however, we choose this method as it is not entirely feasible to detect these changes with the available data, and as under this observation window we are likely to have bias induced by the window size.

Results

We use the KM estimator to find the first two central moments of our sample IETs. For this particular statistic, we limit the analysis so that there are at least four calls, which generate a minimum of three IETs, reducing the number of ties from ~ 26 million to ~ 11 million. Figure 11 displays the joint distributions of $\bar{\tau}_{ij}$ and O_{ij}, as well as the interaction with w_{ij}. Our results show that, the longer the expected IET, the more likely that overlap is smaller, which makes intuitive sense. We find that Pearson$(\bar{\tau}_{ij}, O_{ij}) = -0.153$ and Spearman$(\bar{\tau}_{ij}, O_{ij}) = -0.182$, suggesting a mild effect. In addition, the expected IET is highly intertwined with w_{ij} and the spread of interactions in the observation window: for us to observe large inter-call times within a finite observation window, the number of calls must be reduced. Indeed, on the right plot of Figure 11 we see that our distribution of $Rank(\bar{\tau}_{ij})$ is bounded on the upper right. In general, we find that the mean IET for links with at least four calls is 7.8 days, with a standard deviation of 8.4 days. For links with at least 20 calls, these numbers are 2.4 and 1.7, respectively.

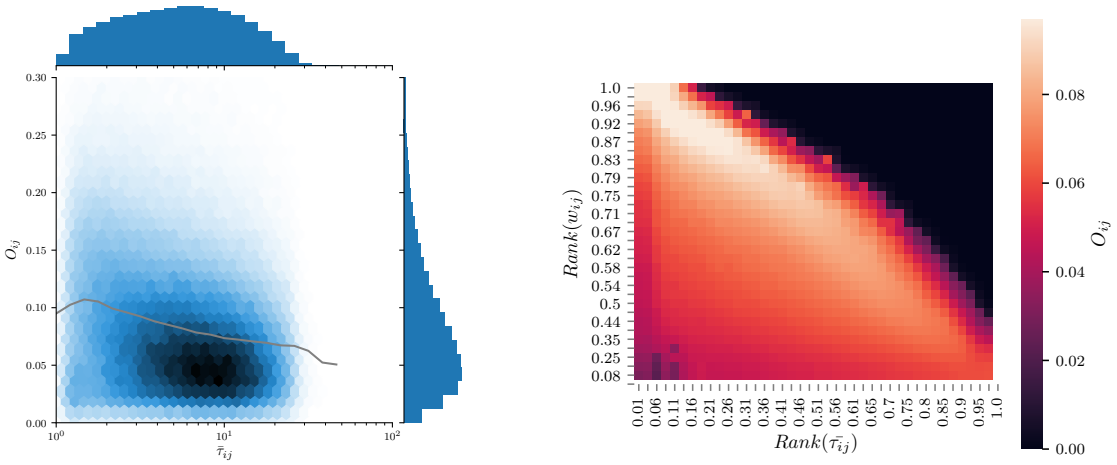


Figure 11: Relationship between expected IET $\bar{\tau}_{ij}$ and overlap O_{ij}, including (*left*) the joint empirical distribution with along with the average overlap , $\langle O_{ij}|\bar{\tau}_{ij}\rangle$ and (*right*) average overlap in terms of both $\bar{\tau}_{ij}$ and w_{ij}.

Figure 12 depicts the standard deviation of the IET distribution using the KM estimator. Indeed, the effect seems to be fairly similar to the expected IET, although for different reasons. In this case, a smaller $\sigma_{\tau_{ij}}$ value could be interpreted as

inter-calls times that are more regularly spaced; as we will see in the next section, however, this might not be necessarily true, as $\sigma_{\tau_{ij}}$ and $\bar{\tau_{ij}}$ might interplay in non-trivial ways through bursty behaviour. For this variables our measures of linear and rank correlation are: Pearson$(\sigma_{\tau_{ij}}, O_{ij}) = -0.157$ and Spearman$(\sigma_{\tau_{ij}}, O_{ij}) = -0.156$, slightly smaller than our the mean IET time.

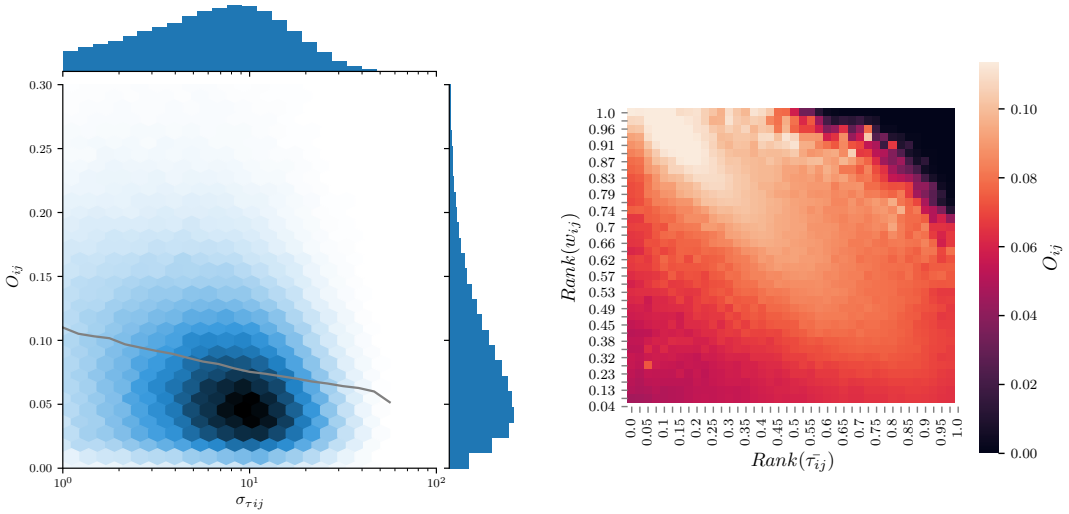


Figure 12: Relationship between IET standard deviation $\sigma_{\tau_{ij}}$ and overlap O_{ij}, including (*left*) the joint empirical distribution with along with the average overlap , $\langle O_{ij} | \sigma_{\tau_{ij}} \rangle$ and (*right*) average overlap in terms of both $\sigma_{\tau_{ij}}$ and w_{ij}.

3.2 Burstiness and Non-Poissonian Dynamics

Human activity is not random; in fact, it displays a *bursty* behaviour, where several actions are performed within a short time, followed by longer periods of inactivity [13, 21, 17]. This behaviour has been observed in various human and natural contexts, from sequences of calls between people and library loans [40] to earthquakes and neural firings [13]. As we previously mentioned (Section 3.1), random activity is usually described as a homogeneous Poisson Process in statistical terms. Nevertheless, communication between two people -and bursty processes in general, cannot be described in such terms.

There has been substantial research into the causes of burstiness in communication networks, with possible explanations varying from circadian rhythms and daily or weekly patterns [20], to human task-execution mechanisms [41] or tie-level dynamics

[22]. In the first case, circadian and weekly patterns have a noticeable effect on call times, since individuals can only place calls when they are awake, and we could expect there to be differences between working hours, nights and weekends, for instance. [20] found that these patterns indeed play a role in burstiness; however, they also noticed that bursty behaviour remained even after removing daily and weekly patterns via a de-seasoning technique, which leads to the conclusion that other factors come into play. Last, [22] concluded that burstiness appears mostly at tie-level, by comparing how the direction of calls and the distribution on ego networks.

Burstiness is a multi-faceted phenomena, an thus admits many characterizations. In this thesis we will focus on three distinct characterizations of bursty processes: one based on the inter-event time distribution, and two based on correlations between IETs. This distinction stems from the marginalization of the IET distribution with respect to time, so that given a series of bursty events, any shuffling of the IETs would yield the same IET distribution. On the other hand, bursty events are correlated, and can thus be interpreted as a memory process [13].

Burstiness coefficient

First, the burstiness coefficient B is strictly based on the first two moments of the IET distribution, $\bar{\tau}$ and σ_τ [13],

$$B = \frac{\sigma_\tau - \bar{\tau}}{\sigma_\tau + \bar{\tau}} \tag{6}$$

Thus defined, B takes values on the interval $[-1, 1]$, and the resulting coefficient may be interpreted in terms of how Poisson-like, bursty or regular the IET distribution is. For the Poissonian case, we expect $\sigma_\tau = \bar{\tau}$, so that $B = 0$; in the event of completely regular intervals, we expect $\bar{\tau}$ to be a constant and for variation to be minimal $\sigma_\tau = 0$, so that $B = -1$; and last, in a bursty case, we expect the standard deviation to be much larger than the average IET, so that $\sigma_\tau >> \bar{\tau}$ and $B = 1$. A similar approach, adopted for instance by [35], is to use the coefficient of variation $\sigma_\tau/\bar{\tau}$, although we prefer definition 6 since its values are bounded. Note, this definition requires for the first and second moments of the IET distribution to exist, which is always the case with empirical data.

The burstiness coefficient B has some drawbacks, and there have been recent

efforts to improve some of its deficiencies [24]. In particular, it has been highly susceptible to finite-size effects, and to be in fact upper-bounded by the number of events in record.

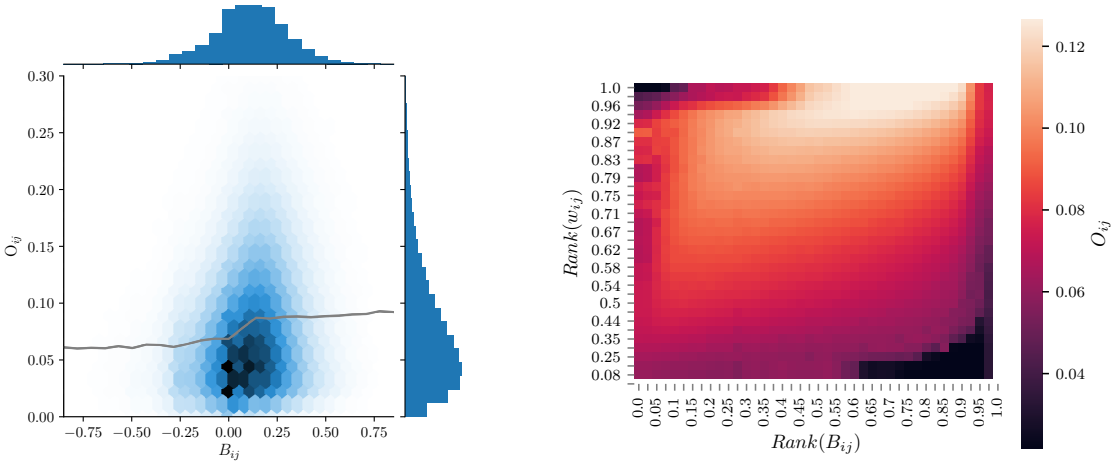


Figure 13: Relationship between burstiness coefficient B_{ij} and overlap. (*left*) Joint density distributions of overlap and burstiness coefficient, with marginalized histograms. The grey line depicts $\langle O_{ij}|B_{ij}\rangle$ showing an increasing trend. (*right*) Overlap as a function of the ranks of B_{ij} and communication intensity, w_{ij}

Figure 13 depicts the behaviour of the burstiness coefficient B_{ij} and O_{ij}. First, note that B_{ij} is positive for 70% of the ties, although still mostly concentrated on the $[-0.27, 0.51]$ interval, meaning that many ties concentrate around zero. There is, however, an increasing trend for the average overlap in terms of burstiness B, and indeed overlap shows its greatest increase only after $B = 0$. When comparing between w_{ij} and B_{ij}, however, it appears that the effect of B is non-linear. The coefficient B gains relevance as it helps characterize high-intensity calls: burstier calls patterns have higher overlap. Nonetheless, for lower call intensity values, less burstiness implies a higher overlap. Although these results suggest more non-linear dynamics, our marginal correlations show $\text{Pearson}(\sigma_{\tau_{ij}}, O_{ij}) = -0.144$ and $\text{Spearman}(\sigma_{\tau_{ij}}, O_{ij}) = -0.155$.

Average Relay Time

The Relay Time τ_R, also known as waiting time, is a concept useful in analyzing spreading phenomena, where we focus on modelling the not the inter-event time, but the time elapsed between a random event time (or a random infection for epidemic models) and the next event. Conceptually, it is derived from the IET distribution, where its distribution may be written as [25]

$$P(\tau_R) = \frac{1}{\bar{\tau}} \int_{\tau_R}^{\infty} P(\tau) d\tau \tag{7}$$

The average Relay Time can be described in terms of the first two moments of the IET distribution

$$\bar{\tau}_R = \int_0^{\infty} \tau_R P(\tau_R) d\tau_R = \frac{1}{2} \frac{\bar{\tau^2}}{\bar{\tau}} \tag{8}$$

Just as the burstiness coefficient B, the average of the Relay Time distribution, $\bar{\tau}_R$, is defined in terms of the first two moments of the distribution, and thus requires similar assumptions, such as a lack of correlation between IETs and finite first and second moments -which is always empirically true, although not necessarily so from a theoretical perspective. In fact, the relationship between $\bar{\tau}_R$ and B is even more powerful, they can be written in terms of one another, as shown in [25].

$$\bar{\tau}_R = \frac{B^2 + 1}{(B-1)^2} \tag{9}$$

Figure 14 displays the relationship between overlap and the average relay time. This relationship is, in fact, slightly easier to understand than the burstiness coefficient. Indeed, it makes sense that if we expect to wait little between a random "infection" and a communication event, then this signifies a stronger relationship. This characterization, however, seems to be better suited for identifying values in the smaller half of the distribution, as the relationship seems to dissipate for larger vales. Now, in this case it seems that the the pair $(w_{ij}, \bar{\tau}_R)$ give a better characterization of overlap than any by itself. All these characterizations derived from the IET distribution have similar linear and rank correlation coefficients, with $\text{Pearson}(\sigma_{\tau_{ij}}, O_{ij}) = -0.154$ and $\text{Spearman}(\bar{\tau_{ij}}, O_{ij}) = -0.154$.

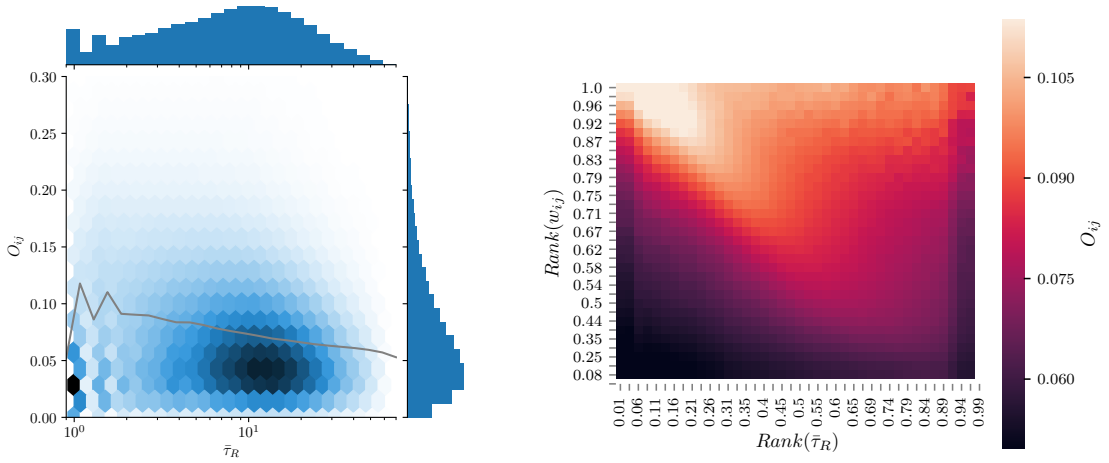


Figure 14: Relationship between average relay time $\bar{\tau}_R$ and overlap. (*left*) Joint density distributions of overlap and burstiness coefficient, with marginalized histograms. The grey line depicts $\langle O_{ij}|\bar{\tau}_R\rangle$ showing an increasing trend. (*right*) Overlap as a function of the ranks of $\bar{\tau}_R$ and communication intensity, w_{ij}

3.2.1 Memory Processes and Bursty Trains

A different approach to modelling burstiness is to understand the generating process as one with memory. Under this scenario, we expect short IETs to follow each other (that is, to have *bursts*), and to observe sporadic calls with large IETs. Although related, memory and burstiness might occur separately, thus it is possible to have a bursty sequence with little to no memory, and vice-versa [13]. In fact, it has been found that by ignoring temporal correlations between IETs, the burstiness coefficient B does not fully capture the way in which bursty phenomena occurs [21]. We will therefore adopt two ways to measure memory effects in bursty sequences: one using a memory coefficient, and another one based on the the distribution of calls *within* a burst, which we will now refer to as bursty cascades or trains.

Memory measures how consecutive inter-event times are related to each other, while B measures the variation of the IET-distribution, irrespective of order. A common way of measuring memory is via the correlation coefficient with a unitary lag [13, 19],

$$M = \frac{1}{n_\tau - 1} \sum_{i=1}^{n_\tau - 1} \frac{(\tau_i - \bar{\tau}_1)(\tau_{i+1} - \bar{\tau}_2)}{\sigma_1 \sigma_2} \tag{10}$$

Where n_τ is the number of events, and $\bar{\tau}_1$ ($\bar{\tau}_2$) is the average IET of the first (last) $n_\tau - 1$ events. This coefficient has been the subject of recent studies, particularly in determining how it affects the auto-correlation function for arbitrary IET distributions [19].

Figure 15 displays our results for the memory coefficient, which has a highly non-linear trend: values concentrated around $M = 0$, where there is little evidence of memory effects, have higher overlap. There is, however, a relevant asymmetry in the rate overlap decay, as it seems that values with positive memory decay more slowly as they depart from zero.

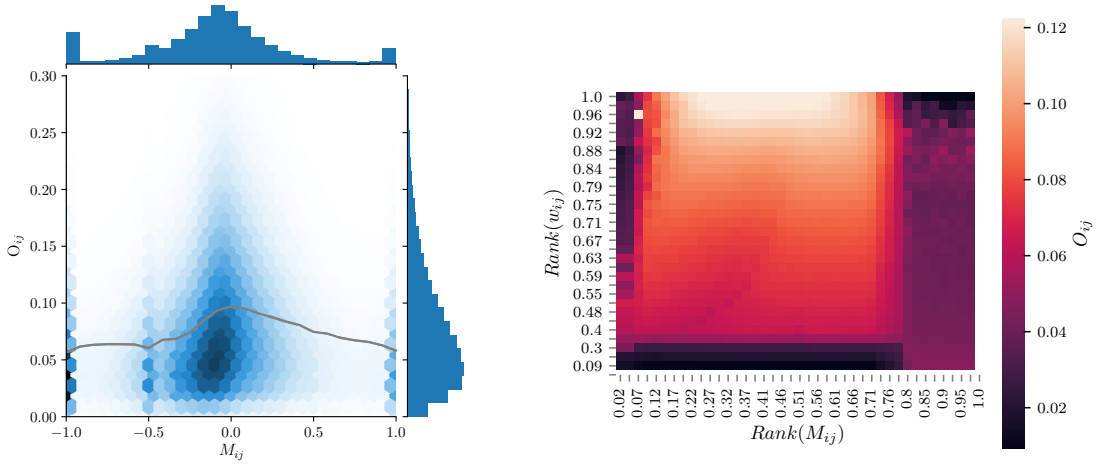


Figure 15: Relationship between memory coefficient M_{ij} and overlap. (*left*) Joint density distributions of overlap and memory coefficient, with marginalized histograms. The grey line depicts $\langle O_{ij}|M_{ij}\rangle$ showing a non-linear trend. (*right*) Overlap as a function of the ranks of M_{ij} and communication intensity, w_{ij}

Our next approach attempts to tackle problems found in measures derived from the IET distribution by focusing on the bursts themselves and what they look like.

Measures derived from bursty cascades

The last set of variables is based on bursty cascades, the distributions of events that are followed within a short time interval Δt [21]. By counting the number of events E within a bursty period, the authors [21] showed that $P(E)$ characterizes temporal correlations not found in neither the IET distribution nor the autocorrelation

function. In short, they found that assuming independence between events, the resulting distribution shows exponential decay, or $P(E = n) \propto \alpha^{n-1}$, for α obtained from the IET distribution; in contrast, empirical data displayed a heavy-tailed behaviour $P(E) \propto E^{-\beta}$ implying that the bursty cascades are decidedly longer (that is, have more calls) than the short bursts induced by an independent model [21], and are thus interpreted as a memory process: there are temporal correlations that the IET distribution itself does not capture.

We may derive many variables from this idea. In this case, we will use N_{ij}^{E}, the number of bursty cascades, $\bar{E}_{ij}$, the average number of events in a bursty cascade and $CV_{E_{ij}}$, the coefficient of variation of events in the busty cascade, estimated as

$$CV_{E_{ij}} = \frac{\sigma_{E_{ij}}}{\bar{E}_{ij}} \tag{11}$$

Where both $\bar{E}$ and σ_E are the emprical mean and standard deviation the number of events.

The use of bursty trains introduces a parameter into the analysis, Δt. The authors [21] found that $P(E)$ is robust for a varied number of window sizes Δt, depending on the context of the phenomena. In the case of mobile phone calls, the authors found no significant difference by using windows of length ranging from seconds up to an hour, although using a Δt of the length of a week might be useless for a wide array of ties [21]. We used $\Delta t = 3600$ seconds (one hour) to analyze our data. For each link we first identified bursty cascades as sequential calls where each consecutive pair had an IET of at most Δt, and recorded the number of events in each bursty cascades.

Figure 16 depicts the relationship between the number of bursty trains and topological overlap. In many ways, this variable is related to call intensity, yet it penalizes all calls within a short timespan and considers them as one. As such, the Granovetter effect for N_{ij}^{E} exhibits similar characteristics as that of w_{ij}, such as a sharp positive increase followed by decreasing overlap - yet the effect seems to be much smaller. In fact, when analyzing both N_{ij}^{E} and w_{ij}, we get an interesting result: ties with a large number of bursty trains have a high overlap, but ties with high call intensity may have a low number of bursty trains, and lower overlap. The correlation between N_{ij}^{E} and overlap is, in fact, the highest of all our variables: $\text{Pearson}(N_{ij}^{E}, O_{ij}) = 0.351$ and $\text{Spearman}(N_{ij}^{E}, O_{ij}) = 0.440$.

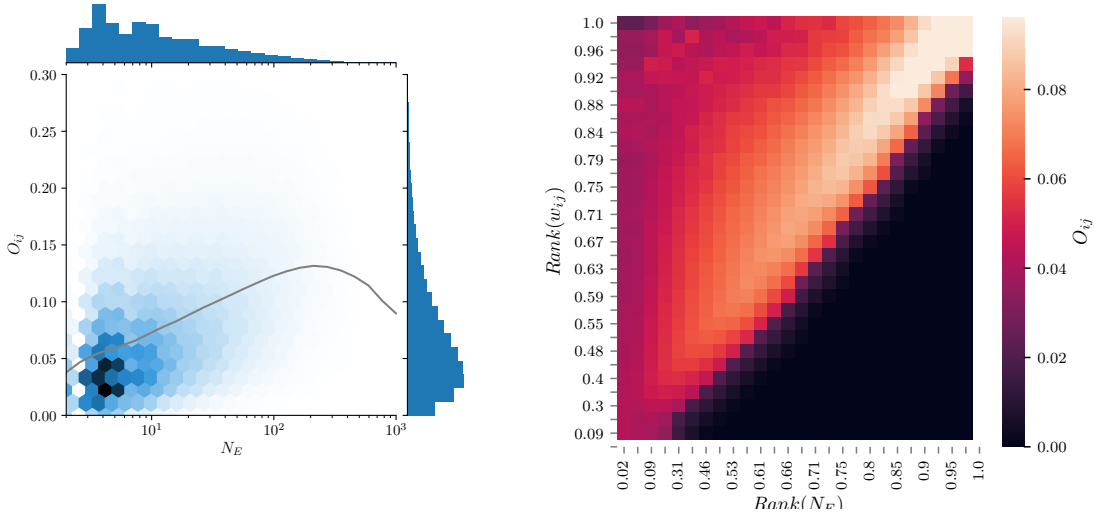


Figure 16: Relationship between average relay time $\bar{\tau}_R$ and overlap. (*left*) Joint density distributions of overlap and burstiness coefficient, with marginalized histograms. The grey line depicts $\langle O_{ij}|\bar{\tau}_R\rangle$ showing an increasing trend. (*right*) Overlap as a function of the ranks of $\bar{\tau}_R$ and communication intensity, w_{ij}

Next, we consider $\bar{E}$, the expected number of events in a bursty train. On Figure 17 we depict the relationship between the inverse $1/\bar{E}$ and topological overlap; we choose the inverse for visualization reasons, as most values are concentrated on near $\bar{E} = 1$, while still being highly dispersed. Our results show some positive correlation, with Pearson$(\bar{\tau_{ij}}, O_{ij}|w_{ij} > 2) = 0.062$ and rank correlation is Spearman$(\bar{\tau_{ij}}, O_{ij}|w_{ij} > 2) = 0.064$, where we require at least three calls to avoid capturing the effect of cases where $\bar{\tau}_{ij} = 1$ merely because few calls were placed.

Last, we shift our focus to the coefficient of variation of events in a bursty train, CV_E. As we have seen, the average number of events might not properly characterize the strength of a tie, yet focusing on the diversity of values that E might take could be more informative. This, however, seems to be only marginally true, as our correlation coefficients improve only slightly, Pearson$(CV_{E_{ij}}, O_{ij}|CV_{E_{ij}} > 0) = 0.078$ and Spearman$(CV_{E_{ij}}, O_{ij}|CV_{E_{ij}} > 0) = 0.089$.

3.3 Measures of temporal stability

In this section we will cover additional temporal features that can be interpreted as measuring the temporal stability of events -not only by how the times between calls

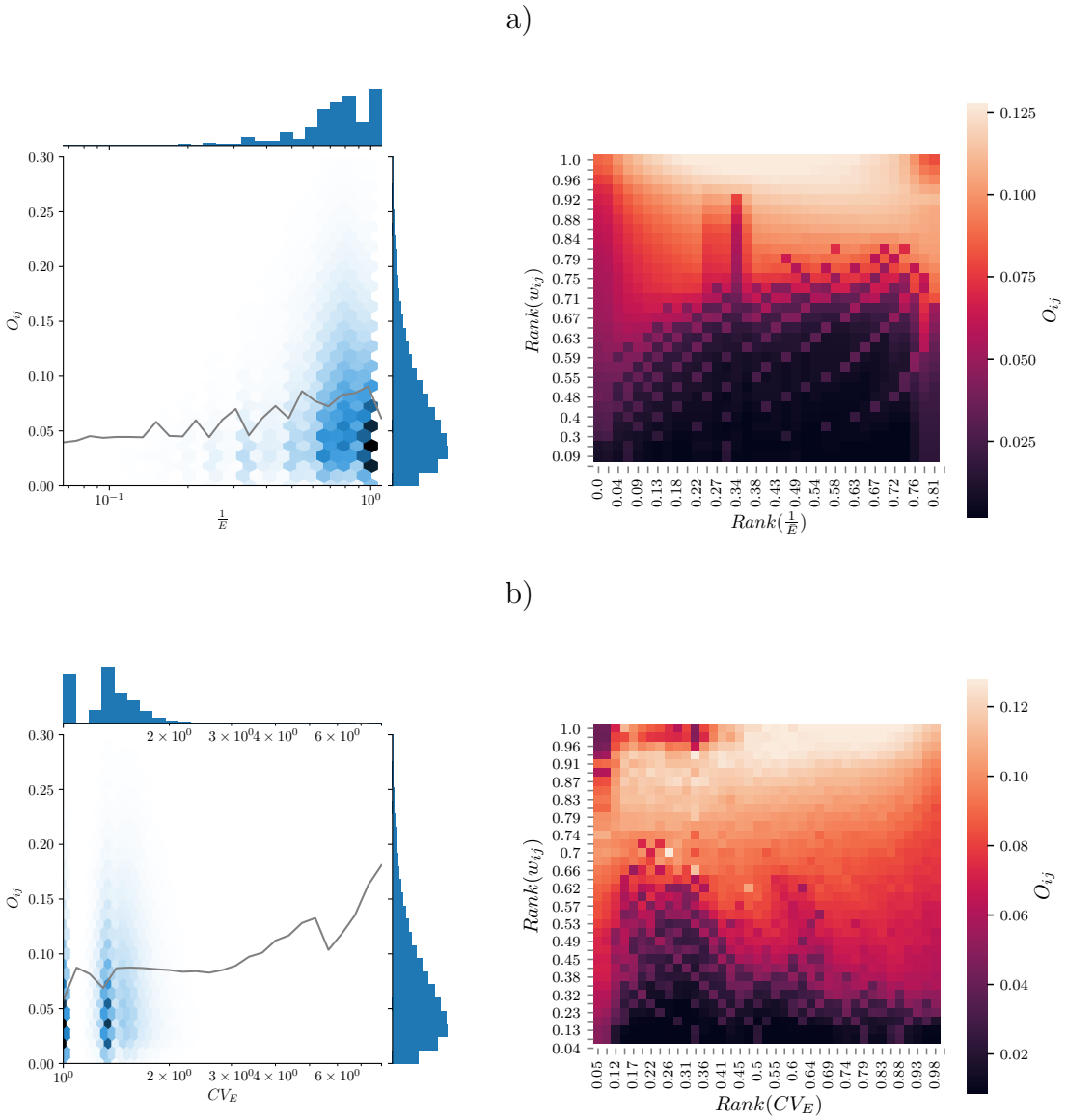


Figure 17: Measures derived from the distribution of events in a bursty train. *a)* Relationship between (inverse) expected number of events in a bursty train $1/\bar{E}_{ij}$ and overlap. (*left*) Joint density distributions of overlap and inverse average number of events in a bursty train, with marginalized histograms. The grey line depicts $\langle O_{ij}|1/\bar{E}_{ij}\rangle$ showing an increasing trend. (*right*) Overlap as a function of the ranks of $1/\bar{E}$ and communication intensity, w_{ij}. *b)* Relationship between overlap and the coefficient of variation - scaled for visualization purposes. (*left*) Joint density distributions and (*right*) comparison between ranks of CV_E and w and overlap.

are distributed, but by *when* the communication events happen in the observation window. Indeed, some of the variables used here are related to the IET-distribution; however, the underlying objective is different, as here our main concern is to obtain

basic features of our calls taking into account the observation window. We will first define three variables to analyze the temporal stability of the interactions within this time window, and propose three other variables that, to the best of our knowledge, have not been used before. These first variables divide the observation window into three: the time before the first event, the time within the events, and the time between the last event and the end of the observation period, as depicted on Figure 18.

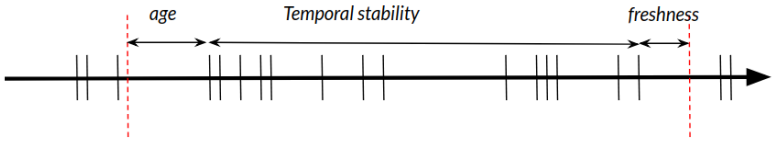


Figure 18: Graphic illustration of variables of temporal stability on the observation window. Although related to the IET-distribution, these concepts attempt to highlight activity within the observation window, and any dependencies there might be.

In the first case we will use a *freshness* variable [39], defined as the time elapsed between the last call and the end of the observation window. Since we attempt to measure whether this last period is larger than average, we use *relative freshness*, which is freshness divided by the average inter-event time $f_r = \frac{T^{end}-t_n}{\bar{\tau}^e}$, where T^{end} is the end of the observation window, t_n the time of the last event, and $\bar{\tau}^e$ the empirical average IET, obtained as the time difference between consecutive observed events. The reason for the use of $\bar{\tau}^e$ and not $\bar{\tau}^{KM}$ is to avoid a biased relative freshness estimate, as $\bar{\tau}^{KM}$ includes information of the last waiting time $T^{end} - t_n$.

In a recent study, [35] found relative freshness to be the most important predictor of tie decay out of a myriad of topological, temporal and user-driven variables. They concluded that if we have already observed a waiting time eight times as large as the average IET, then ties are highly likely to show no activity at all in further observation periods - in other words, that the tie has decayed. As we have mentioned before, it is not entirely possible to define decayed ties in this context; nevertheless, we do know that decaying ties exhibit dramatic changes in local topology, decreasing overlap as shown by Miritello's Dynamical Granovetter Effect [29] which we covered in Section 2.2.2, so the question here will be whether these temporal markers also play a role in non-dynamical contexts.

Figure 19 depicts the relationship between overlap and relative freshness, with a

clear negative trend towards higher values. This results, however, are not comparable to those of [35], as both the scope and objective of our studies differ. In their case, their attempt is to identify ties that will decay during the following observation window. In our case, however, there seems to be a non-linear relationship, particularly when considering the rank of the distribution, as overlap only decreases conditionally on larger f_r values. As previously mentioned, local topology tends to change along with the creation and destruction of ties [29], yet our overlap measurement is for the aggregate network. As a hypothesis, we could say that if the local topology has changed, then we might expect common neighbors to be active only in a small period of time, effectively diminishing the aggregate overlap as more non-common neighbors appear. This, however, would require a different methodology to isolate the effects of freshness and the temporal changes in topology.

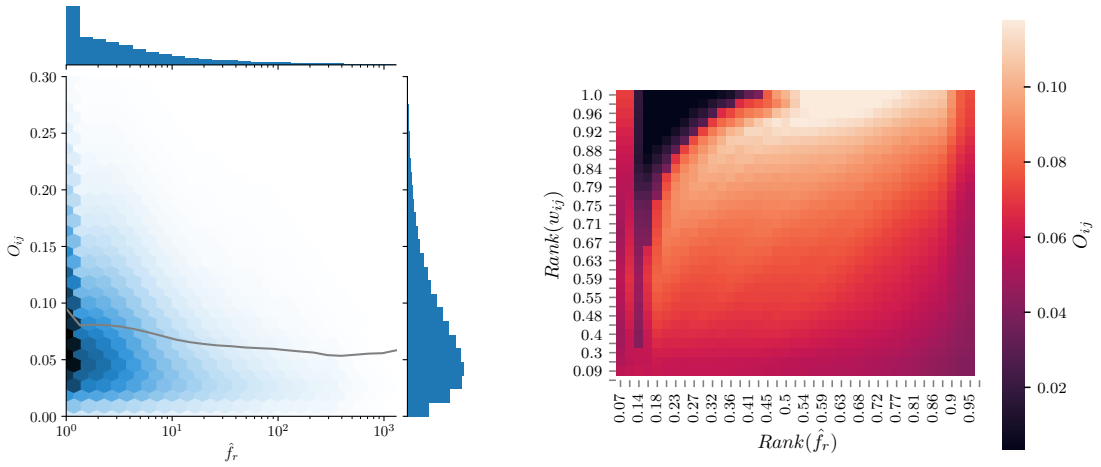


Figure 19: Relationship between relative freshness f_r and overlap. (*left*) Joint density distributions of overlap and relative freshness, with marginalized histograms. (*right*) Overlap as a function of the ranks of f_r and communication intensity, w_{ij}

A variable somewhat analogous to freshness, the *age* [39] of a tie *age* is the time elapsed between the start of the observation window and the first event, $age = t_0 - T^{start}$, for t_0 the time of the first event and T^{start} the starting time of the observation window. We do not divide by $\bar{\tau}$, the average IET because this variable is conceptually different: we expect it to measure the time spent before we observe an event, in a manner more similar to τ_R, the relay time.

We show the distributions of age_{ij} and O_{ij} on Figure 20. The results seem to be promising, as the age of a tie does seem to have an effect in the aggregate topology: average overlap decreases for ties that where first observed at latter periods. This variable also allows us to characterize high-intensity ties that have lower average overlap; namely, intense ties that were "born" at latter periods in our observation window. Interestingly, most of the ties were first observed at the beginning of the observation window. Again, it is not the scope of this thesis to understand the drivers of behaviour behind every variable, yet we hold a possible hypothesis: our data starts on the first of January, so people's calls could be related to ritualistic behaviour stemming from New Year's. This seems to be a sensible reason, particularly since the average relay time distribution is concentrated around $\tau_R \approx 10$ days, larger than the age concentration we observe.

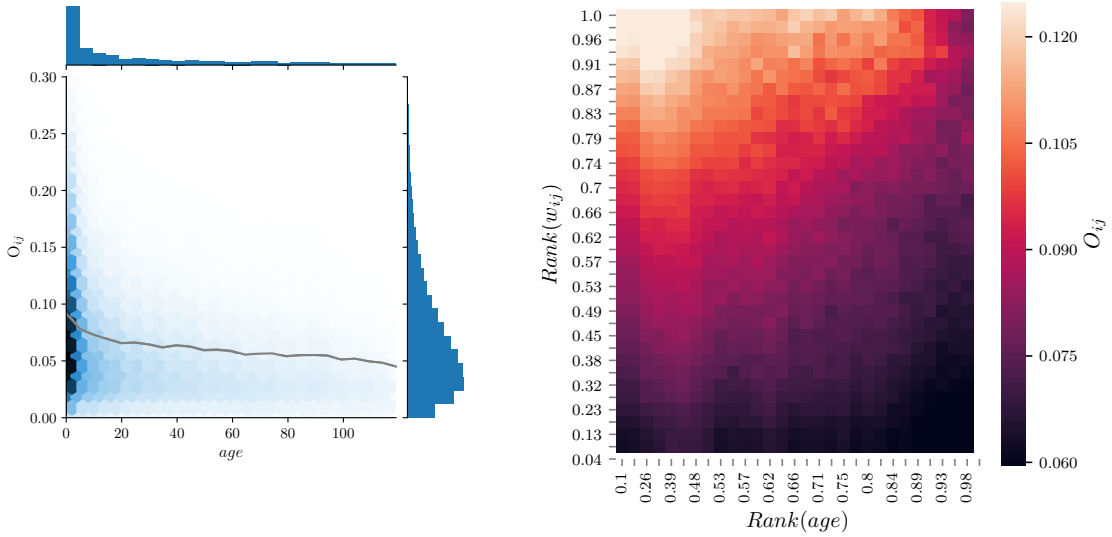


Figure 20: Relationship between the age of a tie age_{ij} and overlap. (*left*) Joint density distributions of overlap tie age, with marginalized histograms, where a large number of ties are (somewhat surprisingly) found within the first few days of the interval. The grey line depicts $\langle O_{ij} | age_{ij} \rangle$ showing slow decay. (*right*) Overlap as a function of the ranks of age_{ij} and communication intensity, w_{ij}

The next variable we will use is temporal stability [31, 39], and defined as the elapsed time between the first and last observed events, $TS = \frac{t_n - t_0}{T}$, where T is the observation window, and t_n (t_0) is the time of the last (first) event. In this case, we would expect a large temporal stability $TS \approx 1$ to indicate that the relationship

lasts at least the length of the observation period, while an extremely short one, $TS \approx 0$, might imply a tie with a shorter lifespan. It bears repeating, the high burstiness of human communication might hinder the reliability of these temporal markers. Additionally, this temporal marker contains no information on intensity, merely informing on the fraction of the period when calls where observed.

Figure 21 confirms our hypothesis that the longer TS_{ij} is correlated to a higher overlap. In addition, this variable provides a useful proxy as it is more evenly distributed during the observation window, whith the exception of with extremely low and extremely high temporal stability, and whose overlap distribution seems to be decidedly dissimilar. It seems that the addition of w_{ij} also provides informative knowledge on overlap, helping distinguish between calls with high temporal stability but low call intensity as those with lower average overlap and calls with high call intensity, whose overlap is higher.

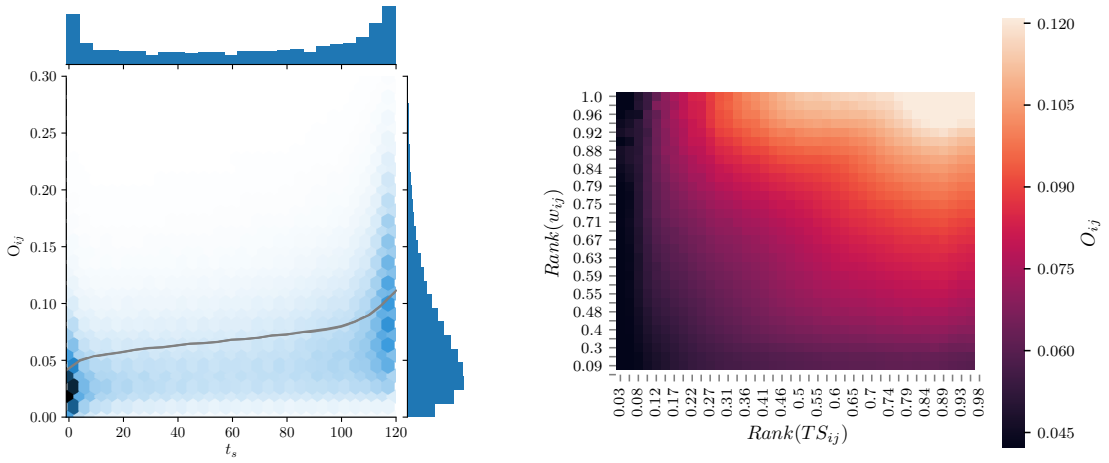


Figure 21: Relationship between temporal stability TS_{ij} and overlap. (*left*) Joint density distributions of overlap and temporal stability, where TS is concentrated on the edges of its distribution, and somewhat uniformly distributed in the central parts. The grey line depicts $\langle O_{ij}|TS_{ij}\rangle$ showing an increasing trend: more stable ties in time are associated to higher overlap. (*right*) Overlap as a function of the ranks of TS_{ij} and communication intensity, w_{ij}, where the use of both variables seems to better characterize overlap.

Additional measures of interaction times

As a last method, we propose a set of variables that attempt to identify *where* most of the mass of interactions is located within the observation window; that is, the average time when we expect the bursty trains to take place. The motivation for these variables has a similar assumption to the previous ones: we expect strong ties to be active independently of the time window, despite the presence of long and bursty inter-event times. Indeed, we opt for using bursty trains as this might ameliorate the effect of several interactions during a short time window. Under this scenario, we wish to examine whether certain temporal bias in the observation window, such as dramatic concentration of interactions at a certain moment in time, might be correlated to a lower or higher tie strength.

To calculate these variables, which we call the average interaction time $\bar{t}^b$, we must first obtain the sequence of bursty trains for a parameter Δt. As we previously mentioned, a bursty train is a sequence of calls that are within a time window Δt of each other. Now, for each bursty train, we define t^b, its time of occurrence, as the time of the first call; note, however, that this might be defined differently based on the whole timespan of the bursty train. Given the sequence of bursty train times of a tie $\{t_0^b, t_1^b, \ldots, t_n^b\}$, we obtain the average $\bar{t}^b = \frac{1}{n+1}\sum_{i=0}^{n+1} t_i^b$. For ease of interpretation, we may normalize each time value t_i^b so that call times are bounded in the unitary interval.

Figure 22 depicts both the average and standard deviation of interaction times, $\bar{t}^b$ and σ_{t^b}, where most of the mass is centered around $\bar{t}^b = 0.5$: for most ties, the average interaction time occurs at the middle of our observation window. This is followed by a trend in overlap, where the highest average overlap also occurs at $\bar{t}^b = 0.5$. Since the relationship between the variables is noticeably non-linar, both correlation coefficients yields results close to zero. For the distribution of σ_{t^b} we have that values are also highly concentrated, now around $\approx .28$, despite being left-skewed. Now, these concentration values do not seem to be random, as the correspond to the mean and standard deviation of a Uniform random variable on the unit interval. This prompted us to devise a new variable that might capture how "uniform-like" the tie's bursty trains are spread on the observation window.

We define a new variable with the following assumption: if there is no observable

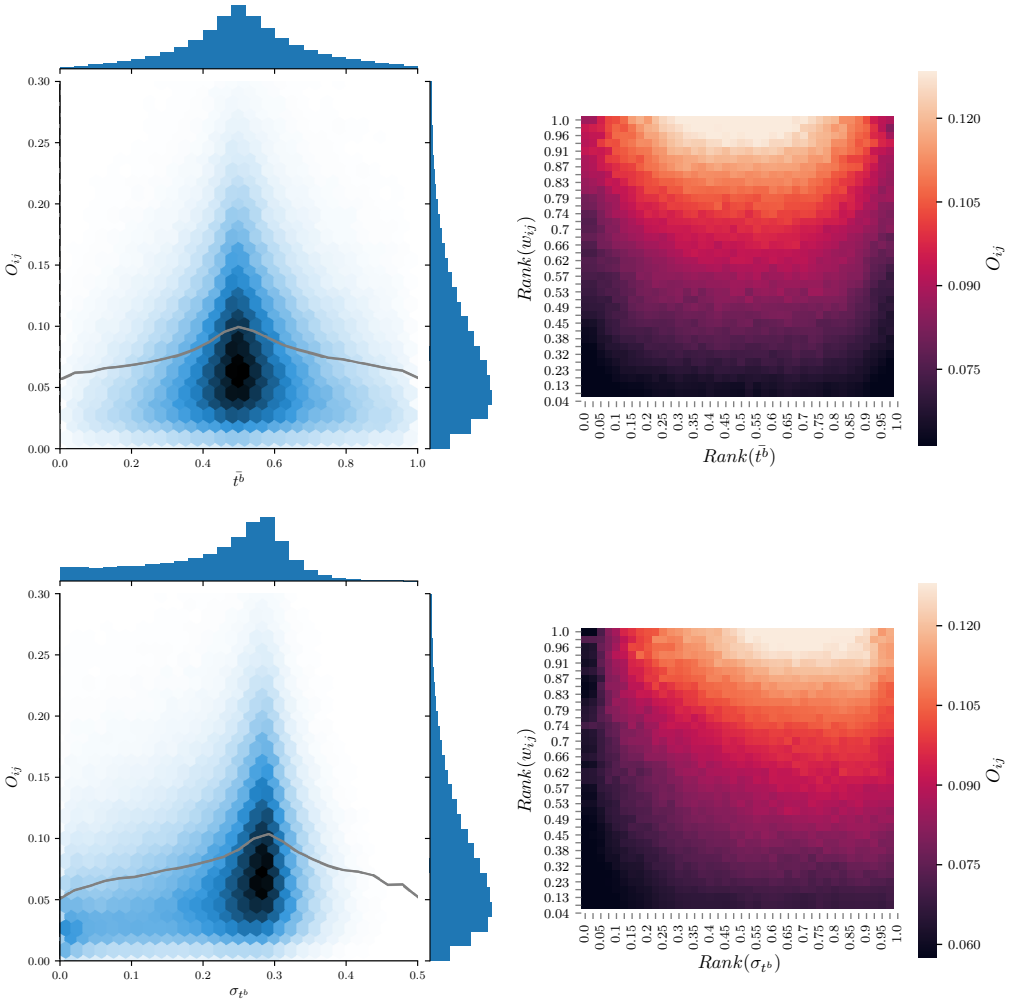


Figure 22: (*top*) Relationship between the average time of bursty trains $\bar{t^b_{ij}}$ and overlap, and (*bottom*) standard deviation of the time of bursty trains $\sigma_{t^b_{ij}}$. (*left*) Joint density distributions, with marginalized histograms, where both $\bar{t^b_{ij}}$ and σ_{t^b} are concentrated on 0.5 and 0.28, respectively. (*right*) Overlap as a function of the ranks of $\bar{t^b_{ij}}$ and σ_{t^b} and communication intensity, w_{ij}. Conditional on call intensity, the main relationship to overlap seems to stem from deviations of the central values.

temporal bias during the observation window, then we expect the average call to be located at its mid-point. We derived a test statistic for difference of means with unknown variance; that is, we build a statistic for the hypothesis test where $H_0 : \theta = \theta_0$; in our case, $\theta = t^b$ and $\theta_0 = 0.5$:

$$T_{ij} = \frac{\bar{t^b_{ij}} - 0.5}{\sigma_{t^b_{ij}} \sqrt{N^E_{ij}}}$$

For this construction, we expect ties where $t^b_{ij} = 0.5$ to have a small value, and ties that differ from the midpoint to have a larger T_{ij} value. Figure 23 contains the relationship between this variable and overlap. Our results suggest that deviations from a central value $t^b_{ij} = 0.5$ indeed show lower average overlap. This variable behaves in a non-linear way: the smallest 60% of values for T_{ij} are non-informative of overlap, conditional on w_{ij}; for the rest of the values, there is a negative correlation between larger test statistics T_{ij} and O_{ij}, which does suggest that a temporal bias within the observation window is correlated to lower overlap.

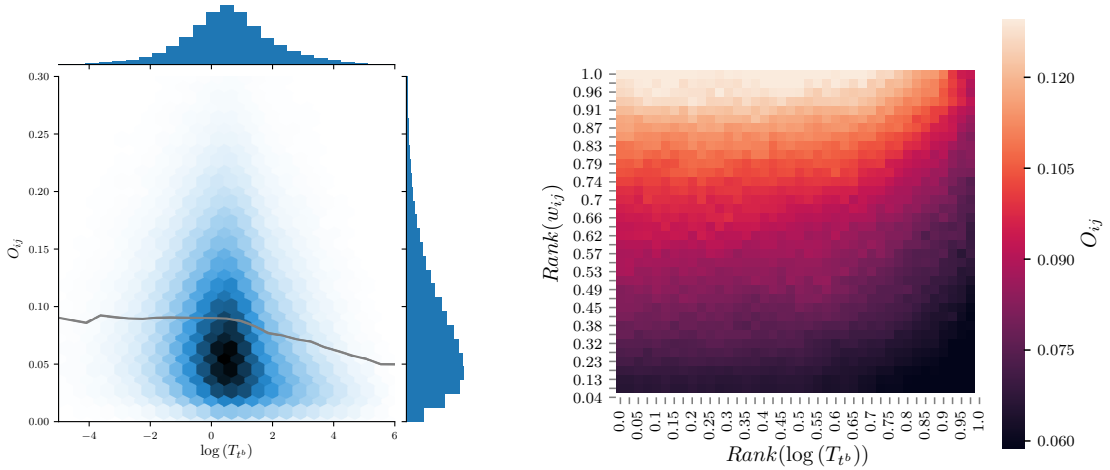


Figure 23: Relationship test statistic T_{ij} and overlap. (*left*) Joint density distributions, where the grey line depicts $\langle O_{ij}|T_{ij}\rangle$ with a decreasing trend for the larger values of T_{ij}. (*right*) Overlap as a function of the ranks of T_{ij} and communication intensity w_{ij}, where T_{ij} is non-informative of overlap for its smaller values, yet shows that large deviations fr $\vec{t}^b_{ij} = 0.5$ decreases the average overlap.

4 Daily and Weekly Patterns

The experience of time is not limited series of events, but we live and behave behave based on cycles. The most natural cycle to think of is the day: all living creatures on earth are bound to 24-hour periods, and we have evolved to adapt to such fluctuations [37]. These cycles are expressed at different levels, as we may think of a biological time, or an inner clock that dictates our daily routines and lives, or a social time, according to which people work, follow schedules and interact with others [37, 2]. In addition, cycles follow different time-scales, where natural cycles such as days or years are intertwined with cultural and social ones, such as weeks, months, or seasonal festivities.

There has been extensive research into the way human activity is synchronized by some of these cycles [37, 45], as well as implications of daily behavioural differences in social networks [4]. For instance, people may exhibit highly heterogeneous communication or activity patterns during the day, according to their *morningness* or *eveningness*, or their propensity to favor certain times of the day [3]. In addition, people's behaviour may be influenced by socially-constructed patterns such as workdays and weekend, which are highly entrenched in cultural, economic and legal frameworks.

In this chapter we will briefly explore how this different understanding of time -not as a linear progression of events, but as a phenomena experienced in cycles-, might manifest in a network's topology. As we will see in the course of this chapter, this does seem to be the case, as people's behaviour at both daily and weekly levels seem to be revealing of the strength of their ties. To see this, we will divide this chapter in two sections according to two different cycle lengths: first the natural daily cycle dictated by the Earth's rotation, followed by the social cycle of seven days that make up a week. In the first case, we the daily activity of people in ties; in the second, we will see whether different time-allocation profiles during the week uncover information about overlap.

4.1 Daily Patterns

Despite all people adhering to daily patterns, each individual may exhibit a particularly distinct activity distribution within the day-night cycle, and may therefore be categorized into distinct *chronotypes*, or groups that reflect the propensity to perform certain activities during different times of the day, such as sleeping, exercising, or performing mental work [1]. Chronotypes have gained prominence in the literature as they seem to be highly informative of individual behaviour and characteristics [45, 1], and usually divided into a *morningness-eveningness* spectrum [1, 3]. In fact, [3] found that evening-types, or people with higher propensity to be inactive during the morning, display homophily in CDR-derived networks. In other words, they found that evening chronotypes tend to communicate more between each other. Measuring chronotypes can be a complicated affair. Although earlier studies focused mostly on questionnaires, during the last decade a new brand of studies has focused on inferring activity from digital sources such as bed sensors and cell-phone activity [34].

In this context, the goal of this section will not be on the detection of chronotypes themselves, but on quantifying differences between the daily time distribution people in ties. This way, we will develop variables in terms of homophily- we wish to see whether similar time-allocating distributions during the day-night cycle are correlated to network topology. We structure this brief section in two parts, first focusing on a framework for measuring differences in distributions, and then developing two variables that summarize these differences.

4.1.1 Measuring differences in daily patterns

In order to quantify how two nodes' behaviour varies from one another, we will define a node's activity to be their outgoing calls. We do not include SMS data, since it has been shown that the same person may have different communication patterns trough different channels [5]. In addition, we do not use SMS data since most nodes have little to no SMS activity. Given the sequence of timestamps of calls for a node's outgoing calls, we categorize each timestamp according to the hour of the day when the call was placed, and finally we divide each hourly count by the total number of calls, to obtain an estimate of the hour-level daily call distribution. In other words, for each person we obtain the parameters of a multinomial distribution,

where each parameter contains the probability of observing a call during one hour, $P_i = (p_0^i, p_1^i, \ldots, p_{23}^i)$. Our goal will be to compare the daily outgoing call distributions of all pairs of people in ties, as well as the differences of tie-level call distribution and people's own distribution.

There are some issues that should be covered regarding the validity of the modelling approach. A first issue refers to the long-term stability of chronotypes, as daily patterns may change during a person's life. These changes, however, are not numerous and tend to occur over large timespans [28]. Another assumption refers to the validity of aggregating the activity of all days into one daily distribution, irrespective of the moment of the week when calls happened. In fact, it is known that people experience working days differently depending on their chronotypes; for instance, evening types tend to suffer from sleep deprivation during the working days because they need to adjust for workplace schedules [34]. This effect can be so large that it may be used to identify evening-types by measuring differences in sleeping schedules between working and rest days [34].

To quantify differences in daily distributions, we will use the Jensen-Shannon divergence (JSD), which is measure of distance between two probability distributions. This approach has been used, for instance, by [5, 43] to show that individuals have persistent daily rhythms that differ from others

$$JSD(P_1, P_2) = H\left(\frac{1}{2}P_1 + \frac{1}{2}P_2\right) - \frac{1}{2}\left[H(P_1) - H(P_2)\right] \tag{12}$$

Where P_1 and P_2 are our empirical probability distributions, and H is the Shannon entropy, an information-theoretic measure commonly used because of it's ability to handle events with zero probability [5]. If P_i is a discrete probability function over time t, then $p_i(t)$ for each time interval

$$H(P) = -\sum_t p(t)\log(p(t)) \tag{13}$$

4.1.2 Results

For each node, we counted the number of outgoing calls placed on each of 24 one-hour bins, and divided for the total number of outgoing calls, using the CDRs with non-

company users to guarantee a full daily profile for each person's calls. In summary, for each person we have a daily distribution $P_i = (p_0^i, p_1^i, \ldots, p_{23}^i)$, where $\sum_{t=0}^{23} p_t^i = 1$.

We denote $JSD_{ij} = JSD(P_i, P_j)$, the JSD for two nodes in a tie, which we refer to as *divergence of daily patterns*. Figure 24 depicts the relationship between JSD_{ij} and overlap, with and without w_{ij}. Results suggest a strong negative relationship with a non-linear factor on the edges of the distribution. All in all, it seems that smaller divergence is correlated to smaller overlap; in other words, that the more similar the daily activity two nodes is, the larger their topological embeddedness. For the case $\langle O_{ij} | Rank(JSD_{ij}, Rank(w_{ij}) \rangle$, the relationship with JSD_{ij} seems to be strong, as the lower ranks of JSD are decidedly associated with a higher overlap, enabling us to identify, for instance, ties with a high call intensity but low average overlap.

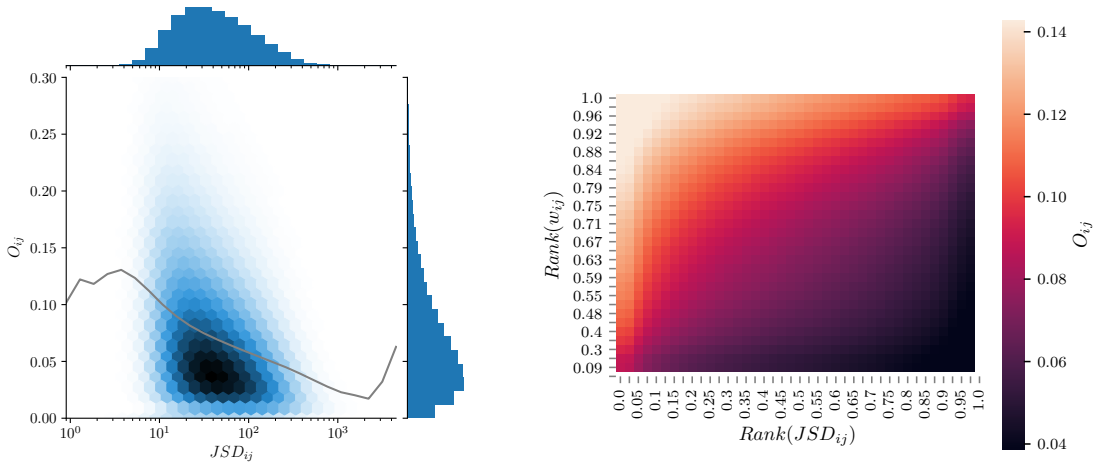


Figure 24: Relationship between divergence of daily patterns JSD_{ij} and overlap. (*left*) Joint density distributions of overlap and difference in daily patterns, with marginalized histograms. The grey line depicts $\langle O_{ij} | JSD_{ij} \rangle$ showing a decreasing trend. (*right*) Overlap as a function of the ranks of JSD_{ij} and communication intensity, w_{ij}.

These results suggest behavioural tie-level homophily, where people with similar daily patterns seem to be embedded in communities, however, this might raise more questions than offer immediate answers, particularly since the relationship between O_{ij} and JSD_{ij} might not be straightforward to uncover. First of all, as a note of caution, these results do not let us conclude that similar behaviour implies a community structure around any two nodes, but that once we have observed a

tie, their daily patterns might be informative of their relationship. There are some explanations as to why we think this might occur: first, other type of latent homophily could drive these results, such as people being from the same age group (for instance, teenagers or college-age students who have similar schedules). Another reason might be related to tie-level rituals, such as family members who call each other at specific times of the day, or friends who know when their friends are available for calling. Last, some people might engage in reciprocal bursty cascades, placing several calls during the same bins, a hypothesis more powerful if considered in conjunction to the two previous cases or if the two nodes spend most of their time calling each other.

Although understanding the dynamics that explain these results is beyond the scope of this thesis, we will explore another case use of JSD and daily patterns. Namely, we will compare how the daily call pattern of a person compares with the daily pattern of all calls with another node. In other words, we will use the sequence of calls of a tie and create a daily tie-level distribution: $P_{ij} = (p_0^{ij}, p_1^{ij}, \ldots, p_{23}^{ij})$ and compare it with P_i and P_j, defining $JSD_{i\to j} = JSD(P_i, P_{ij})$. Now, this variable is biased by construction, particularly since calls are counted twice in the construction of p_t^{ij} and p_t^i. We justify our approach, however, since we believe this measure might help us understand whether the call profile of a link fits well into a person's daily pattern, or whether the way a person communicates with another is, in a way, anomalous in her daily distribution. Now, this leads us to two variables per link $JSD_{i\to j}$ and $JSD_{j\to i}$. We compute the difference of daily patterns to links $JSD_{ij}^{\text{diff}} = |JSD_{i\to j} - JSD_{j\to i}|$, which we depict on Figure 25. This variable may be interpreted in terms of asymmetry of a tie, as large values will occur if one person's daily pattern is substantially dissimilar from the tie-level behaviour. Results show a decreasing relationship between overlap and JSD_{ij}^{diff}, including in the presence of communication intensity w_{ij}.

4.2 Weekly Patterns

We will now switch our focus to see whether the time of interaction during the week is related to network topology. There are many reasons to do so, as we expect people's activities to vary at different points of the week [2]. We may, for instance, hypothesize that people communicate more with their loved ones during weekends or

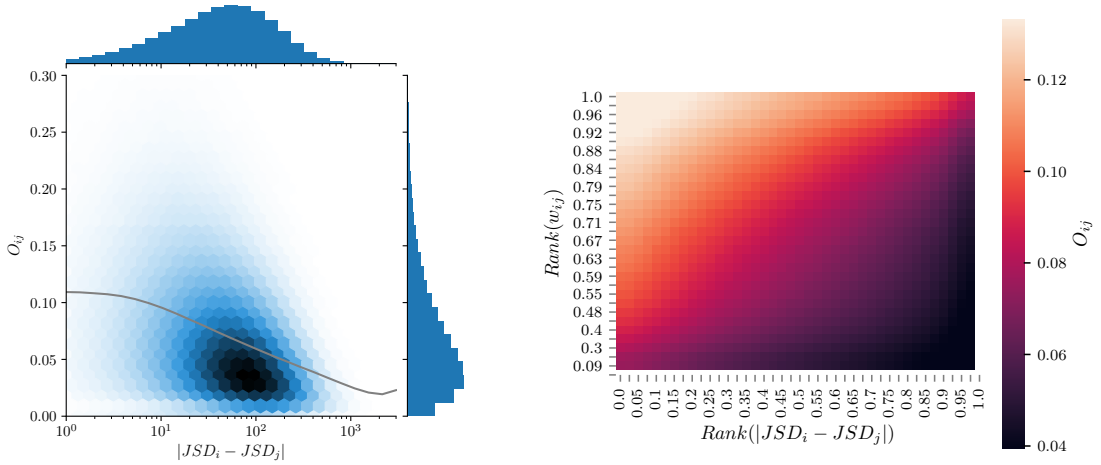


Figure 25: Relationship between the difference of daily patterns to links JSD_{ij}^{diff} and overlap. (*left*) Joint density distributions of overlap and burstiness coefficient, with marginalized histograms. The grey line depicts $\langle O_{ij}|JSD_{ij}^{\text{diff}}\rangle$ showing a decreasing trend. (*right*) Overlap as a function of the ranks of JSD_{ij}^{diff} and communication intensity, w_{ij}

during non-working hours, while maintaining other forms of communication at other times. In other words, our approach here will differ from the previous section as we will now focus on identifying whether there are any specific communication times that are informative of topological structures.

The task at hand is to obtain variables that serve as profiles for people's weekly activity, and then use these variables to understand overlap. There are many possible solutions to this problem. One way would by to use a top-down approach, where we bin the distribution of based on set of rules set by the researches (for example, saying that all hours between 09:00 to 12:00 during working days belong to the same category). Instead, we will a high-granularity approach to divide the week into hourly bins -meaning that for each link we will have a weekly profile $7 * 24 = 168$ hour-based variables- as done by [2], and then we will work bottom-up to define informative clusters of hours based on the way our population places calls.

We proceed as follows: first, we will do exploratory analysis of the weekly call distribution, where we will examine the aggregate activity of our population during the week, as well as correlations on the distributions of times. Then, based on some results from the exploratory analysis, we will form clusters of hours and evaluate them based on their ability to explain differences in overlap values.

4.2.1 Exploratory Analysis

We divide the week into 168 hourly slots, corresponding to an hour-based binning. For each tie, we will then count the number of interactions that happen per hour and obtain a weekly normalized distribution of communication events, which we denote as the set $\{\phi_{ij}^h\}_{h=1}^{168}$, where ϕ_{ij}^h is the proportion of calls placed between nodes i and j during hour h of the week. For most links the majority of hourly bins will be empty, as most ties have a small number of calls which are then distributed among a small number of bins. Our objective will be to identify which groups of hours are associated to a higher overlap, taking into account that many of these bins are empty. We include the restriction $w_{ij} > 4$ in order to guarantee that we focus on ties that are more likely to exhibit patterns.

We start our exploratory analysis with Figure 26, where we plot three time-series of aggregate statistics from our data. There are clear weekly patterns that emerge not only in terms of communication intensity, but also on the relationship to overlap. First, there is a clear bi-modal daily distributions that is mostly consistent through weekdays, but changes shapes for Friday, Saturday and Sunday. Correlation to overlap *(b)* also shows a bi-modal daily distribution that peaks on Saturday afternoons. This effect, however could still be explained in terms of communication intensity, which in previous sections we disassociated from w_{ij} by plotting overlap in terms of both $Rank(w_{ij})$ and $Rank(V_{ij})$ for ther variables V_{ij}. Since it is not possible to see for a set of 168 variables, we now define decoupled overlap:

$$O_{ij}^{-w} = \frac{O_{ij}}{\langle O_{ij} | w_{ij} = w \rangle} \tag{14}$$

Where $\langle O_{ij} | w_{ij} = w \rangle$ is the average overlap of links with weight w. This new variable takes value $O_{ij}^{-w} = 1$ if it's link overlap is entirely explained by communication intensity, while takes values larger (less) than 1 if the link has a larger (smaller) overlap than that expected by it's weight. We find that certain hours do reflect higher-than expected overlap, particularly for the weekend, which suggests that the *times* when people contact each other are indeed markers of community structures around them.

As a note of caution, we must take into account that this data-set belongs to a

certain country, and thus there are behavioural patterns that are cultural. Certainly, different countries have different customs and legal frameworks regarding working days, and people's behaviour is tightly related to them. The behaviour we obserb is thus very context-specific.

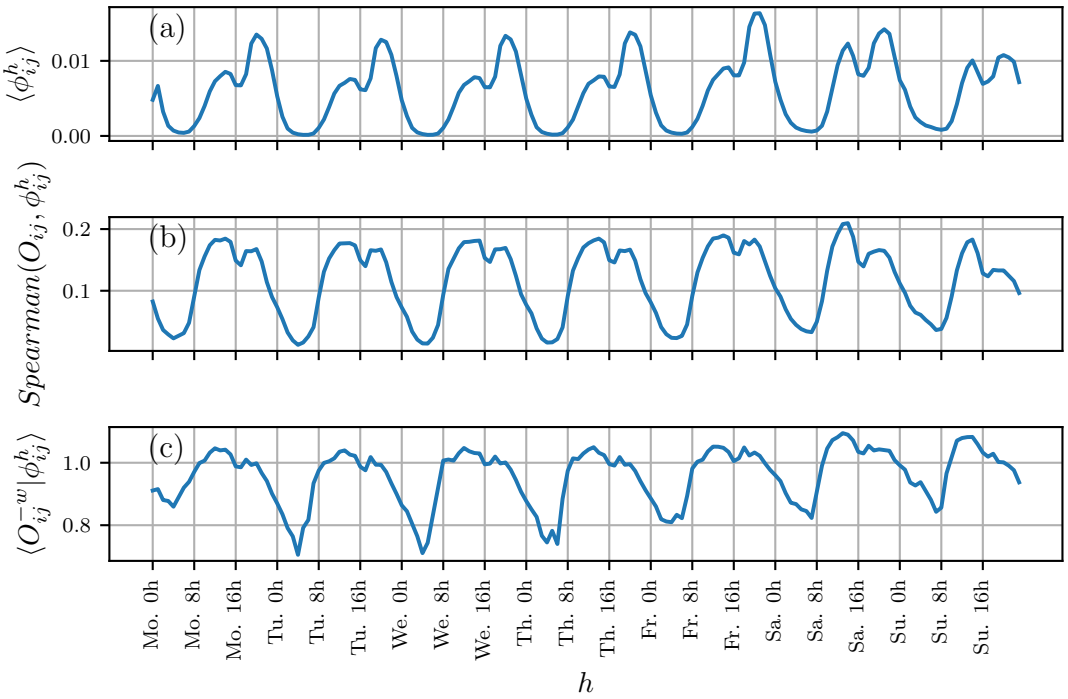


Figure 26: Network statistics per hourly bin h. *(a)* Average hourly profile value $\langle \phi_{ij}^h \rangle$, *(b)* Rank correlation between overlap and hourly profiles $Spearman(O_{ij}, \phi_{ij}^h)$, and *(c)* Average overlap per hourly profile where *overlap is decoupled for link weight*, $\langle O^{-w} | \phi_{ij}^h \rangle$. Weekends are associated to higher-than expected overlap, as well as weekday noons.

We continue our exploratory analysis by studying the correlation matrix of $\{\phi^h\}_h$, which tells us which hourly bins have similar activity patterns, depicted on Figure 27. The matrix displays clear structures on workday, day, and hour-levels, so that it is possible to visually discern between different days. Indeed, most activity is performed roughly from around 7:00 *am* to 1:00 *am* of the following day, and the low-activity times create a grid of values with little to no correlation. Daily patterns play a substantial role, and are visible in the form of equally-spaced titled lines of varying intensity. Indeed, these lines may be interpreted as similar activity patterns at the same hours of different days. Noticeably, they decrease in intensity during the last days of the week, suggesting that people place calls differently during the

workdays and weekends.

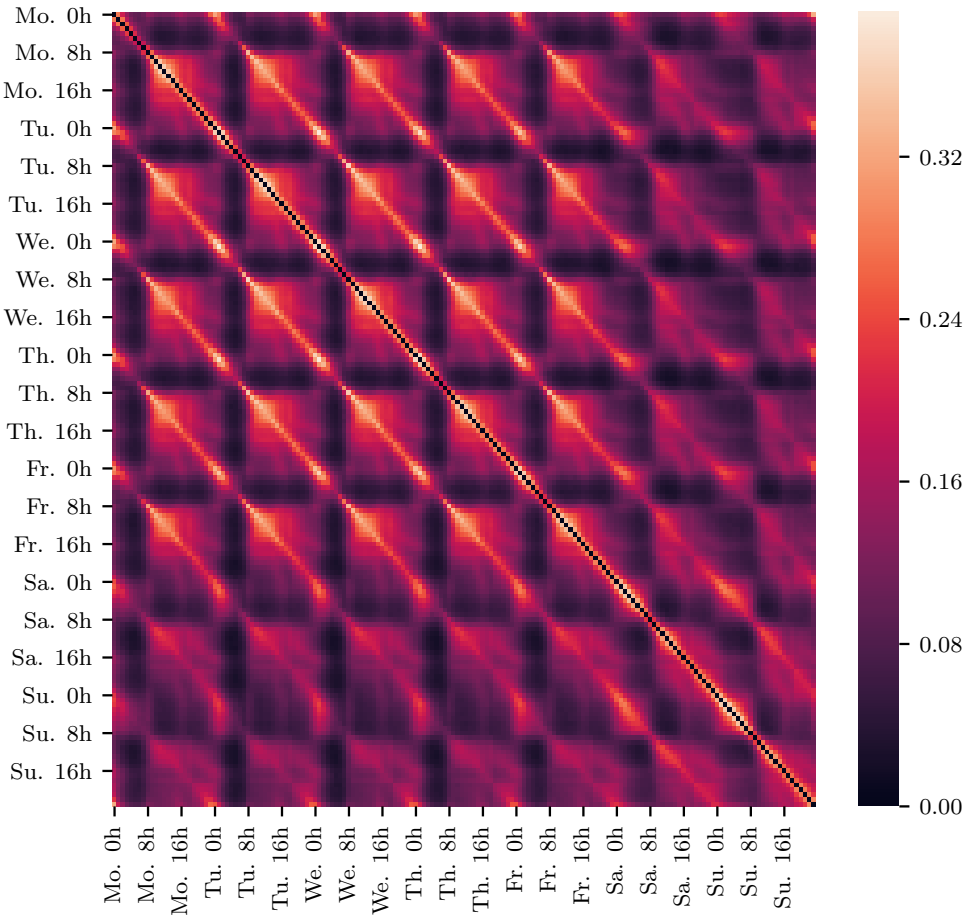


Figure 27: Correlations for hourly bins during the week, where the diagonal values (equal to 1) are not shown for visualization purposes. The correlation matrix shows several clear structures: first, most hours of the working days are correlated to the same hours on different days. In addition, there are strong correlations near the diagonal, meaning that hours are mostly related to the hours of the same day. Weekends are clearly dissimilar from the rest of the week, favouring correlations near the diagonal (on the same day, as opposed to similar hours on other days).

Next, we performed principal component analysis (PCA) on the weekly profiles $\{\phi_{ij}^h\}$. In short, PCA is a method used in dimensionality reduction that projects correlated sets of variables onto orthogonal subspaces, thus yielding new, uncorrelated variables called principal components. This way, if $\{\phi^1, \ldots, \phi^n\}$ is the set of $n = 168$ variables from the weekly bins, PCA will produce n new variables of the form $PC_i = \alpha_{i,1}\phi^1 + \alpha_{i,2}\phi^2 + \ldots + \alpha_{i,n}\phi^n$, where PC_i's are linearly independent of each

other [16]. PCA may be derived as an optimization problem that finds a linear combination of variables that maximize the amount of explained variance from the original data; as such, each PC_i will explain a smaller amount of variance. The resulting PCs correspond to the eigenvectors of the correlation matrix.

We analyzed the variance explained per principal component for two cases: when the data obtained is from $\{\phi_{ij}^h\}_h$, the link's hourly profile, and $\{I_{\phi_{ij}^h>0}\}_h$, where each hourly bin is assigned the value 1 if at least one call was placed during that bin. In this sense, while the hourly profiles $\{\phi_{ij}^h\}_h$ are more akin to a probability distribution over a week; the binary profile places the same value to all times when any activity was present. The resulting PCA decompositions capture wildly different levels of variance: the first component of the weekly profile captures 2% of the variance, while for the binary profile the first component captures 21% of the variance. Indeed, for the binary weekly profiles, the first principal components capture a high percentage, only to decrease rapidly. On the other hand, the PCs for weekly profiles have a consistently low variance explained, meaning that a large number of the variables are needed to explain the data variation: while people *place* calls in somewhat predictable times, they *distribute* calls in these times in highly varied manners, which cannot be easily captured in a small number of dimensions.

We now inspect the weights that each hourly bin has on the principal components; that is, the $\alpha_{k,h}$ values for PC k and bin h. Figure 28 depicts the weights for the first five principal components, which in total account for 9.7% of the variance for the weekly profiles, and 31% of the variance of the binary profiles. We may interpret the coefficients of the PCs as a decomposition of the signal, informing of different ways in which variation occurs. To do so, we focus on both the sign and magnitude of the $\alpha_{k,h}$'s: similar weights in a PC imply that these hours have a similar importance in the component and they vary at similar rates. In contrast, different signs for two variables imply that a certain amount of variation happens separately. The first PC of the binary profile (orange) has variable coefficients that mimic the hourly average activity, meaning that most variability can be explained by assigning similar weights to *important* hours. On the other hand, the first PC for the weekly profile $\{\phi_{ij}^h\}_h$ (blue) separates behaviour for weekdays and weekends. Indeed, weekdays have mostly negative coefficients, while Friday and Saturday evening have highly positive

spikes. This may be interpreted as signaling that we may distinguish among pairs of people that distribute their calls either during the weekdays or during weekend evenings. The second PC adds nuance to this asseveration, assigning positive values to weekday nights, including Fridays.

We may summarize some of the takeaways from the first PC coefficients:

- The behaviour from Monday to Thursday is mostly similar, with hourly patterns being more relevant than the specific day of the week
- Weekends are not easy to generalize this way, as neither Saturday nor Sunday have consistent patterns between themselves
- Friday displays a mixture of weekday and weekend behaviour, being more similar to weekdays in the morning, are more similar to Saturday nights in the evenings.

In summary, the exploratory analysis has allowed us to uncover a series of relationships between weekly temporal patterns and topological structure. In addition, we have found that although people place calls at highly predictable hours, the way they distribute these calls is heterogeneous. The use of weekly bins for overlap prediction might be counterproductive, however, as these variables are rather sparse for most observations. For this reason, we will now perform dimensionality reduction on the bins to see wheteher it is possible to retain information of overlap while using larger temporal resolutions than hours in a week.

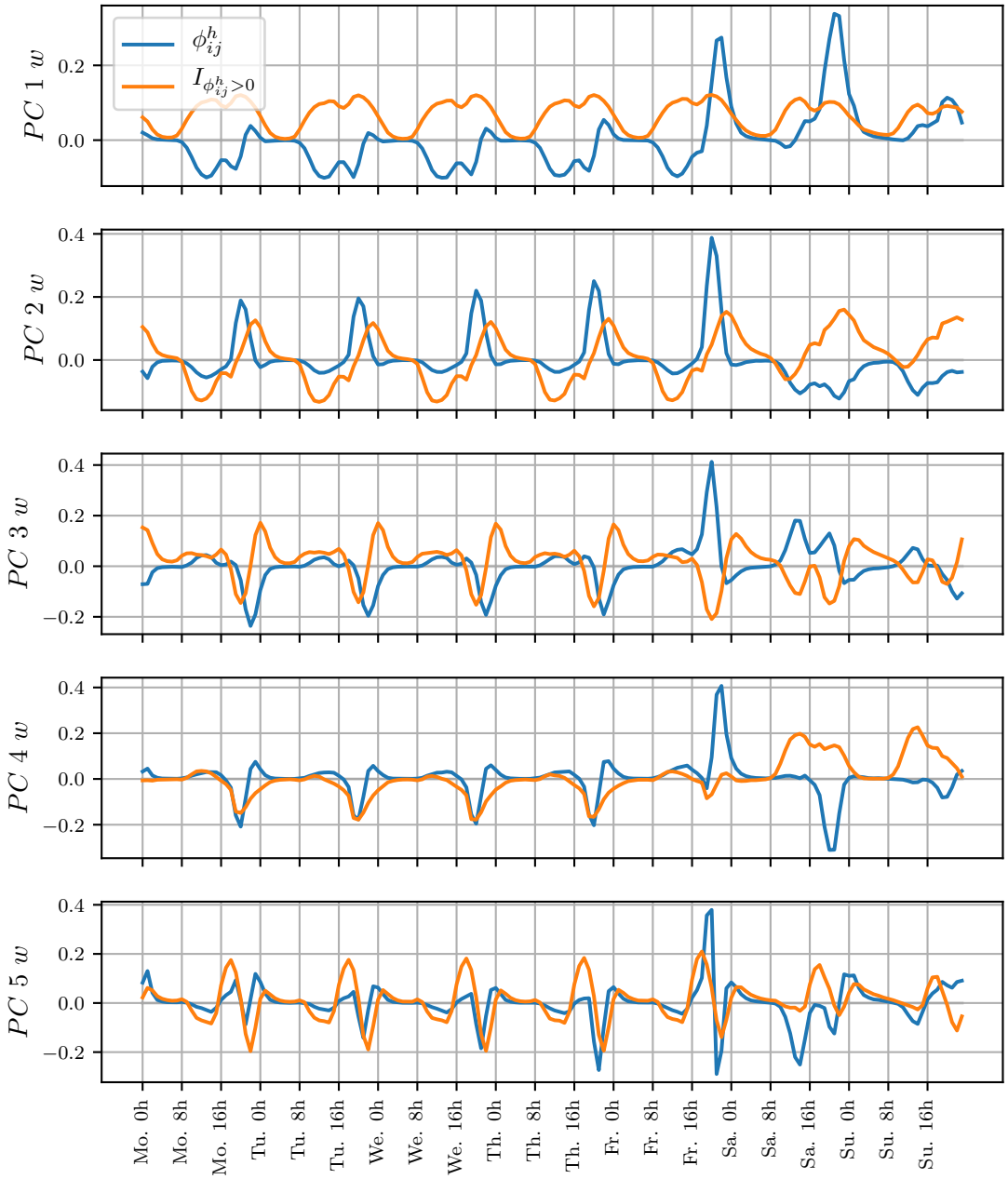


Figure 28: Weight of each hour on the first five principal components for weekly distribution. (*blue*) profile of weekly hour communication, ϕ^h_{ij}, and binary profile of communication, $I_{\phi^h_{ij}>0}$.

4.2.2 Dimensionality Reduction

In this section our goal will be to reduce the number of variables while maintaining valuable information about overlap. In short, we wish to perform dimensionality reduction on the variables of the weekly profiles $\{\phi_{ij}^h\}_h$, while retaining the interpretability of our variables in terms of the fraction of calls placed in those hours. PCA, one of the most common dimensionality reduction technique, yields variables that are not easy to interpret in terms of the original variables, particularly for higher dimensions, and in our case, do not seem to contain substantial information about overlap (for example, the first PC has $Spearman(PC_1^{\phi}, O_{ij}) < 0.03$).

There is a wide array of approaches that we may use for dimensionality reduction. Here, we propose a methodology based on two different objectives: first, clustering bins that might capture similar activity based on the times on which people call each other; second, finding the sets of clusters that explain more variation of overlap together. The rest of this chapter is therefore used to explain our approach to dimensionality reduction, using Markov Cluster Algorithm (MCL) [46] for generating new variables, as well as their evaluation in terms of overlap.

Markov Cluster Algorithm

The Markov Cluster Algorithm (MCL) is a methodology developed for clustering graphs by [46]. It is based on the idea that if there is a random walk on a graph, then it will likely visit many of the vertices of a cluster before leaving it. Although developed for finding clusters in graphs (in other words, communities of observations), we adapt it to be used for variables by generating graphs from the correlation matrix. This way, we use the same source of information that PCA uses, although we gain more control over how to aggregate our variables.

In a simplified manner, we may think of MCL as a methodology that simulates a random walk on a graph in matrix form, where first we simulate a walk -in a step call expansion-, and then we strengthen (weaken) the stronger (weaker) links by taking powers of weights between nodes -a step called inflation- [46]. After some steps, natural communities begin to appear between the variables. For more information on the algorithm and reasoning behind it consult [46]. We choose this methodology for a series of reasons. First, it is a fast and scalable method that does not require for us

to specify the number of clusters beforehand - the algorithm determines the natural cuts itself. Second, it focuses on capturing clusters by identifying groups where all elements are highly connected among themselves; that is, it does not sequentially add or remove elements. This becomes relevant as it takes into account how communities behave as a whole. Other approaches that we experimented with involved the use of other clustering algorithms (such as K-means), or proposing a minimization problem, yet the clusters yielded by MCL seemed to outperform these other methods.

As a dimensionality reduction methodology we follow the next procedure: given the correlation matrix, we go over different cutoff values ψ that determine a graph, establishing a link between two bins if their correlation is higher than the cutoff value ψ. For each ψ we obtain a clusters C_ψ via MCL, and create new clustered weekly profiles $\{\phi^c_{ij}\}_{c\in C_\psi}$ where $\phi^c_{ij} = \sum_{h\in c}\phi^h_{ij}$, thus retaining the interpretation of each cluster as the fraction of calls placed during that cluster. There is a strong positive relationship between ψ and the number of clusters, as a small ψ will imply that most of our bins are connected, resulting in a small number of large "communities". A large ψ implies that simply a few bins are connected, and there is a large number of small "communities". We then evaluate our clusters based on the amount of variation they explain of overlap, and select the smallest ψ value that explains the largest overlap. Note that although we do not provide a theoretical guarantee that such ψ value exists, for this dataset we find that it does.

We use a sample of $200,000$ links with weights $w_{ij} > 4$ to analyze how our bins are related to overlap. We go over different values of $\psi \in [.1, .4]$ to obtain different clusterizations, three of which we depict on Figure 29.

As previously mentioned, our target is to choose a ψ that yields the minimum number of clusters and explains a sufficient amount of overlap variation. We analyze this by evaluating the range of values in overlap that the clusters capture. Given a set of clustered profiles, $\{\phi^c_{ij}\}_{c\in C_\psi}$, we obtain the distribution of weighted average decoupled overlap per cluster, $\{\langle O^{-w}_{ij}|\phi^c_{ij}\rangle\}_{c\in C_\psi}$. In other words, we decouple overlap and weight by obtaining the average overlap value for weights as in equation 14, and then for each cluster c, we obtain the average overlap for those profiles. Figure 30 depicts the distributions of overlap values, $\{\langle O_{ij}|\phi^c_{ij}\rangle\}_{c\in C_\psi}$. For a smaller ψ, clusters are not able to capture large differences of overlap, as most bins are grouped in large

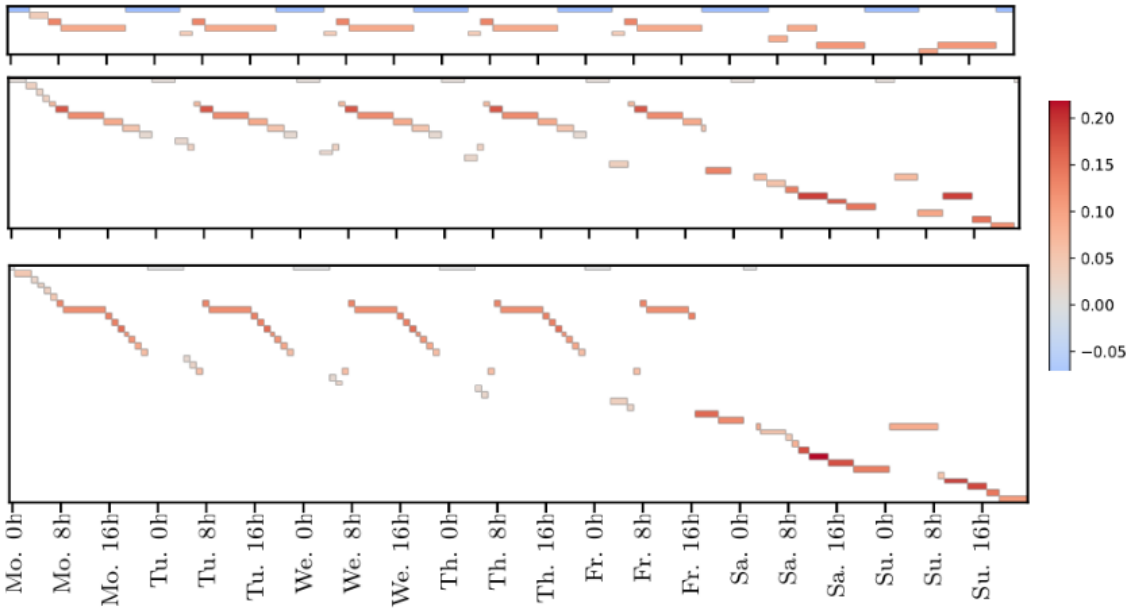


Figure 29: Clusters created using MCL for graphs created with different values of ψ. The x-axis depicts the hour of the week and the y-axis the (unique) cluster it belongs, while the color represents the cluster's Spearman correlation with overlap O_{ij}. (*top*) MCL clusters for $\psi = .18$, resulting in 8 clusters; (*middle*) case for $\psi = .22$ and 25 clusters, and (*bottom*) for $\psi = .255$ and 39 clusters. In all three cases, weekdays' clusterization follows a similar pattern, with weekends requiring a larger number of clusters than the rest the days. A larger cutoff value ψ results in a larger number of clusters, with days of the week being categorized in similar clusters per hour, and weekend days on their own clusters.

clusters and the effect on O_{ij}^{-w} dilutes. Increasing ψ yields larger variation up until a certain degree, around $\psi \in [0.20, 0.3]$. We choose $\psi = .2$ to generate our clusters since this value contains a sufficient amount of diversity in overlap values.

On Figure 31 we display an additional visualization of our chosen clusters to use for the task of overlap prediction. This methodology thus seems to capture blocks of clusters that are adjacent both on hour and day-based scales. Interestingly, behaviour during the weekdays between 3 and 6 am generate a large number of clusters. This might be related to the fact that usually few calls are placed during that times, so it is not so easy to generalize the behaviour of links that place a significant amount of calls at such hours.

Now, although here we focus on how individual clusters explain overlap, these clusters might be used to determine more complex behavioural profiles. For instance, it would be interesting to know whether different time-allocation patters, such as Friday nights against workday mornings has a relationship of network topology. In

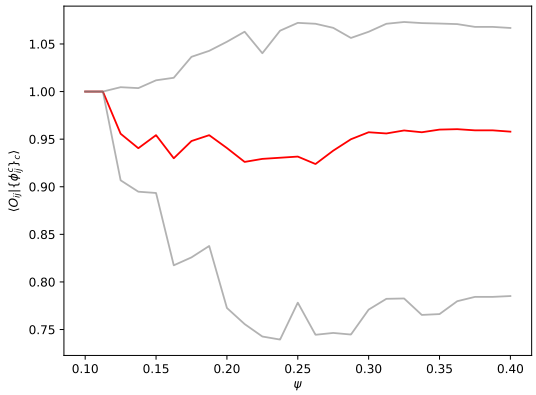


Figure 30: Average decoupled overlap distribution for clusters, $\{\langle O_{ij}^{-w}|\phi_{ij}^{c}\rangle\}_{c\in C_\psi}$, as a function of threshold values ψ. In red, the average overlap of the distribution; in grey, 5% and 95%-percentiles of the distribution. For $\psi = 0.1$, all bins belong to a single cluster, while at $\psi = 0.4$ every bin constitutes its own cluster. The variation of

addition, it is a possibility that different kinds of relationships with high overlap communicate at different times, such could be the case of weekday mornings (cluster $C4$) and weekend afternoons (cluster $C12$), both of which have high mean overlap, but could be qualitatively different.

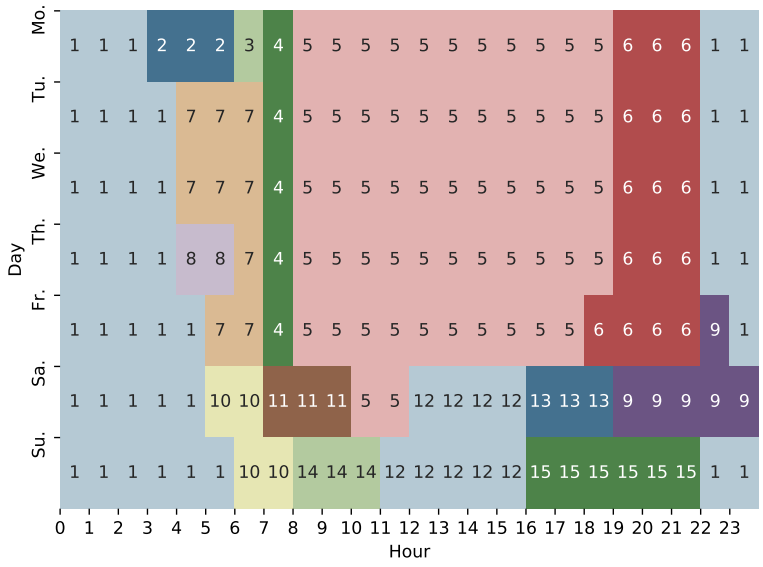


Figure 31: Clusters obtained via MCL for cutoff value $\psi = 0.2$, where each number corresponds to the variable name used in the following chapter. There are clear weekly structures.

5 Prediction of Topological Overlap

In this final chapter we discuss the problem of predicting overlap using the variables we previously examined. Indeed, we believe that both topology and behavioural features are expressions of the strength of a tie, so finding a relationship between our features and overlap might reveal which behavioural markers are important for community structures in social networks. So far we have analyzed sets of static and temporal features, revealing a series of associations between behavioural features and network topology that sometimes provides additional information not captured by our communication intensity measure, w_{ij}. That being so, we have mostly omitted discussions about how these variables might be related among themselves, and how they might be used either by themselves or interacting with other variables for predicting overlap.

As before, our focus here will be on being able to interpret our results. Hence, we favour models that allow us to assess the influence of variables, not necessarily models that yield the best predictive performance. We structure this chapter in the following manner: first, we briefly explain the experiments and metrics we use. Since we wish to examine how different variables might yield the most information on network topology, we will conduct a series of tests, where we first train single-variable models, followed by double-variable models that include w_{ij}, and finalize with a full-variable model. In addition, we compare the predictive performance of our features with mean temporal overlap (2.2.4), to see whether similar sets of variables are able to explain overlap as measured in time.

5.1 Models and methods

We follow a strategy that showcases how overlap is related to our behavioural features, using machine learning models in different scenarios by means of different sets of variables: first, we isolate the effect of our features by modelling overlap using every variable as a single predictor. As we have seen, some variables have non-linear relationships to overlap, thus using different ML models allows us to generalize a feature's ability to predict overlap that go beyond linear measures, such as correlation. Indeed, communication intensity w_{ij} might not be an optimal predictor of overlap,

so a second scenario will focus on using pairs of variables as predictors of the form (V_{ij}, w_{ij}), where V_{ij} are the other variables. This second scenario has a dual objective, as we will analyze whether the predictive power of our models increases or decreases when compared to single-variable models, and whether variable V_{ij} is relevant for the dual model via its *feature importance*. This will allow us to gauge whether there are variable combinations more powerful than simply using w_{ij} as a measure of tie strength. A third scenario incorporates all variables for the prediction task, where we analyze again model performance and feature importance. This will allow us to determine which are the most relevant features given the full set of variables. As a last experiment, we repeat the third scenario but we change our target variable for its temporal counterpart, as we attempt to understand whether different variables are useful in determining temporal overlap.

Instead of predicting a specific overlap value, our main focus will be to determine whether it is possible to classify overlap as either high or low determined by a series of cutoff values O_α. Certainly, while modelling overlap as a regression task is valuable in itself, the use of different high/low overlap classifications will allow us to monitor both the range of overlap values that are best suited for prediction and the effect of different features at different overlap levels. In addition, preliminary regression tests suggested that it is not trivial to diminish variance in our predictions. For all these reasons, we choose an approach based on going over different O_α values that cover the whole range of the overlap distribution, varying at 5-percentile intervals.

In all four scenarios, we will use three models commonly used in machine learning tasks: Random Forests (RF), Extra Trees (ET) and Logistic Regression (LR). Here we will briefly describe these models and how to evaluate their performance. The first two cases are ensemble methods based on decision trees; in other words, models that use sets of decision trees in order to obtain weighted predictions [16]. Logistic regression is a statistical model based on the logistic function used for the prediction of binary variables [16].

Decision trees are, intuitively, a series of binary tests, where each test divides the feature space into into two [16]. As such, decision trees explore the subspace of feature relationships by finding splitting the data to best describe the target variable. As mentioned, both RF and ET are based on the aggregate use of decision

trees, where each tree is trained only on a sample of features and data, and the final prediction is obtained via the average of the trees, a technique that reduces over-fitting [16]. The main difference between RF and ET is, in a nutshell, that they use different algorithms for splitting the feature space (that is, for creating the tree). As such, RF compute an optimal split, while ET use random values for splitting.

For these methods, the *importance* of each variable may be estimated using *mean decrease impurity* (MDI), a measure related to the number of times a variable is used in a splitting rule, weighted by the number of samples used in that split [16]. As the name indicates, MDI uses an impurity measure that captures the splitting rule's classification ability, and how much a splitting rule *decreases* the average impurity [16].

LR models follow a different approach, and are a way of specifically modelling binary target variables. Given features $\{x_1, \ldots, x_n\}$, a logistic regression finds coefficients $\{\beta_0, \beta_1, \ldots, \beta_n\}$ such that [16]

$$P(y = 1|x, \beta) = \sigma\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i\right)$$

Where $\sigma(t) = \frac{1}{1+\exp(-t)}$. The importance of a variable x_i for the prediction of y can thus be thought of in terms of its coefficient β_i: positive values increase the probability of $y = 1$, while negative values decrease it. Note, however, that this modelling approach attempts to capture somehow linear relationship among variables, where the probability of $P(y = 1)$ is modelled as increasing/decreasing linearly with each variable x_i. Finally, we cover measures of performance for classification. Given a binary classification problem, we may measure the performance of our model via a confusion matrix:

	Positive	Negative
Predicted positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True negative (TN)

According to a specific problem, one may choose to evaluate a model's performance according to different metrics performed on this confusion matrix [10]. For instance, accuracy (ACC) is defined as the rate of true positives and true negatives among the whole population. This metric, however, may not give information on classification if

the number of positive and negative samples is extremely asymmetrical (for example, if only 5% of our observations are positive, then classifying all the sample as negative yields $ACC = 95\%$). We choose to use the Matthew's Correlation Coefficient (MCC) to evaluate the performance of our models since it takes into account all cells in the confusion matrix, and is more robust to differences in class sizes [10]. The MCC indeed captures correlation between observations and predictions, and so may be interpreted accordingly: $MCC = 1$ implies a perfect prediction, $MCC = 0$ is expected when predictions are not better than random guessees, and $MCC = -1$ occurs when every element is mislabled [10]. In terms of the confusion matrix, MCC is defined as [10]

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

5.2 Results

We selected a random sample of $400,000$ links with at least five calls (out of a total of $5,775,901$ links that met this criteria), and used the full data set of features derived from the CDRs in order to predict binary classifications of topological overlap O_{ij}. For each of the models, we used 3-fold cross validation and report mean values for MCC and feature importance. For each of the four scenarios (single-feature predictors, dual-feature predictors, all-variables predictors and temporal overlap) we ran three models: Random Forests (RF), Extra Trees (ET) and Logistic Regression (LR) for 20 classification tasks corresponding to different cutoff values of O_α

First, we present a summary of our variables on Table 2 for ease of reference, sorted according to their performance on the single-variable models. The clusters names correspond to rough approximations to the hours they contain, with the exact hours presented in the previous chapter on Figure 31.

Figure 32 depicts the MCC for the first two scenarios of our classification tasks, where we sort the variables according to their mean MCC in the single-predictor model.

There are several relevant results in the single-variable scenario. First, this definition of tie strength - via a relationship to overlap, admits several characterizations that do not only depend on intensity measures: the best average predictive perfor-

Symbol	Name	Reference
N^E	Number of bursty trains	3.2.1
JSD_{diff}	Difference in daily patterns to links	4.1
TS	Temporal Stability	3.3
w	Communication intensity	2.3.2
JSD	Daily pattern divergence	4.1
$C12$	Weekend afternoon cluster	4.2
σ_{t^b}	Standard deviation of times of bursty trains	3.3
$C1$	Night and early morning cluster	4.2
$C9$	Saturday night cluster	4.2
$C6$	Weekday night cluster	4.2
$C15$	Sunday evening cluster	4.2
$C5$	Weekday morning and afternoon cluster	4.2
$\bar{E}$	Average events in a bursty train	3.2.1
$\hat{f}$	Relative freshness	3.3
$C13$	Saturday evening cluster	4.2
$\bar{\tau}_R$	Average Relay Time	3.2
age	Age	3.3
σ^E	Standard deviation of events per bursty trains	3.2.1
σ_τ	Standard deviation of IET	3.1.1
M	Memory coefficient	3.2
$\bar{\tau}$	Average IET	3.1.1
CV^E	Coefficient of variation of events in bursty trains	3.2.1
$\bar{t}^b$	Average time of bursty trains	3.3
$\log(T)$	Test statistic for times of bursty trains	3.3
$C14$	Sunday morning cluster	4.2
$C11$	Saturday morning	4.2
B	Burstiness coefficient	4.2
$C4$	Weekday 7 am	4.2
$C10$	Weekend early morning cluster	4.2
$C7$	Tuesday - Friday 5-7 am cluster	4.2
$C2$	Monday early morning cluster	4.2
r	Reciprocity	2.3.2
$C8$	Thursday early morning cluster	4.2
$C3$	Monday 6 am	4.2

Table 2: Summary of temporal features used for overlap prediction, ordered according to the results on Figure 32, and reference section where the feature was introduced.

mance belongs to the number of bursty trains N^E, difference in daily patterns to links, JSD_{diff}, temporal stability TS and communication intensity, w. This confirms that not only are intensity-derived measures related to community structures, but there are several possible characterizations, including communication profiles at specific times. Another important result is that variables have a rather limited useful range of predictive values, misclassifying most observations at extreme values of O_α, and most variables tend to perform best at relatively shorter and central O_α values. Indeed, w seems to have a rather restricted valid range, smaller than that N^E, and TS. On the other hand, differences in daily patterns JSD and JSD_{diff} have large valid ranges, despite not achieving the highest scores for any O_α value. Moreover, some of the least strong predictors gain importance at smaller O_α's, such as relative freshness f_r, memory M, the average time of bursty trains $\bar{t}$, and burstiness B.

For the dual-variable predictors, the inclusion of variable pairs does not seem to increase significantly the performance of most models when compared to w, except for (JSD, w) and (JSD_{diff}, w) in logistic regression. This is slightly surprising, as it seems that the enforced linearity of the model on the variables based on daily patterns has a general positive effect that neither RF or ET models capture. It seems that the greater effect of including variable pairs (V_{ij}, w_{ij}) is expanding the range of O_α values for which predictions are possible. In terms of feature importance, most models favour variables that capture the lower end of the overlap distribution not represented by w. Indeed, variables such as f, M and B have larger feature importance for the dual models, even if their overall performance is not strong. Last, it is noteworthy to mention that both N^E and TS have relatively small feature importances, which is mostly due to high correlation with w, meaning that these sets of variables do not work optimally together.

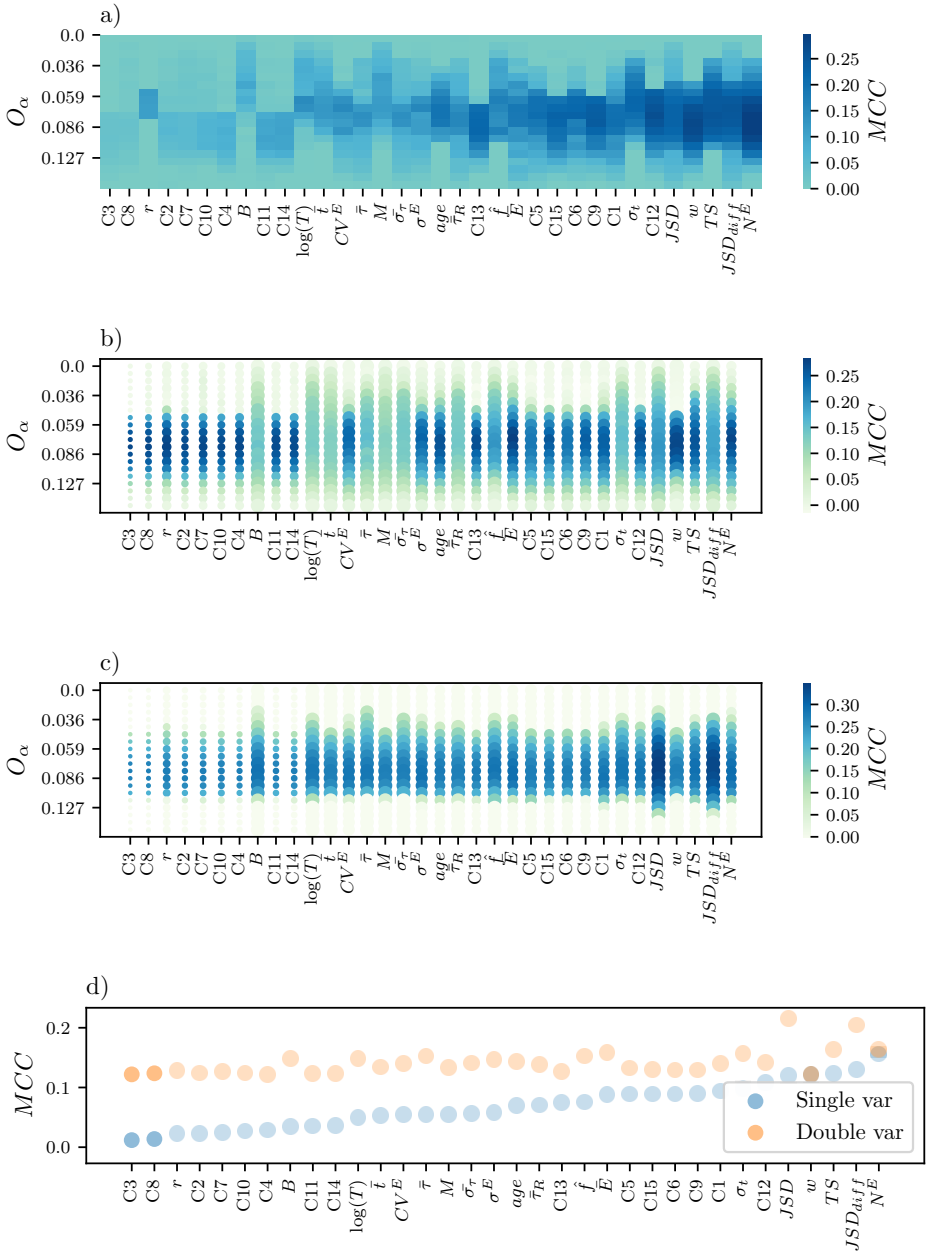


Figure 32: Matthew's Correlation Coefficient (MCC) for overlap prediction in different scenarios, where the x-axis represents variables, the y-axis represents different cutoff values α for binary classification of high/low overlap, and the color represents MCC, a model performance metric. *a)* Average MCC for three models trained with **single-feature predictors**, where each variable is used to predict overlap using RF, ET and LG. *b)* - *c)* MCC (color) and feature importance (point size) using **dual-feature predictors** of the form (V_{ij}, w_{ij}), using models *b)* RF, and *c)* LR, where the size is determined by the LR coefficients. We do not include results for ET as they are virtually identical to RF. *d* Comparison between single and dual-variable models, where we depict the average across all models and cutoff values O_α.

Our third and fourth scenarios, where we use the full set of variables to predict static and temporal overlap, are depicted on Figure 33 (MCC scores) and Figure 34 (feature importances). Let us first focus on the MCC score, which follow distributions that peak for central O_α values. As in the previous two scenarios, this is mostly due to the fact that there is high variability in overlap values, so classification is a difficult task. Indeed, by definition of O_α the high/low overlap sets are asymmetric for extreme O_α values, yet testing the use of weighted sampling to correct for these asymmetries only improved our estimates marginally. Interestingly, RF and ET seem to perform better for extreme values, while LR is by far the best model for central O_α's, so different variables and weights become relevant for different definitions of high/low overlap. We only depict the prediction of temporal overlap $\hat{O}^t_{ij}$, which contains generally better MCC scores; that is, the prediction capacity of these features is indeed higher for all three models, suggesting a stronger relationship between our variables and temporal overlap, with the best model overall being a LR with $-O_\alpha = 0.063$, and overall accuracy $ACC = 73\%$. This might be due to the fact that temporal overlap penalizes "common neighbors" that do not have calls with both nodes i and j during the same period, effectively diminishing the overlap values for the lower end of the distribution. On the other hand, it also maintains high overlap values for cases where "common neighbors" maintain consistent communication through time, which might be a better indicator of tie strength.

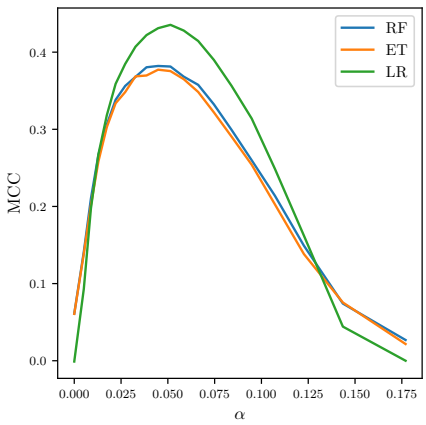


Figure 33: MCC fourth scenario, using the full set of features to predict mean temporal overlap $\hat{O}^t_{ij}$ for twenty cutoff values α. We do not depict the results for the third scenario since it follows a similar distribution, expect that it's peak value if $MCC = .038$ with the LR model.

We now focus on feature importance of our variables for the full models depicted on Figure 34, where the top two graphs depict static overlap O_{ij} and the bottom two depict mean temporal overlap $\hat{O}^t_{ij}$. We may summarize our results in some key findings: first, communication intensity w_{ij} does not play a relevant role in the presence of all the other variables. Indeed, on the LR model for static overlap, it has a negative coefficient, meaning that it is actually being used to penalize overlap. In all models, N^E and JSD play important roles for capturing overlap. For the temporal scenario the mean inter-event time $\bar{\tau}$ and temporal stability TS become prominent for the LR model. Now, there are some notable differences between our ensemble-based models and LR in the importance of variables, particularly for some of the cases that we know do not have a linear relationship to overlap, such as σ_t, $\log(T)$, M and $\bar{t}$. Indeed, these variables seem relevant for our ensemble models, but not much for our LR models, which assume a linear relationship among variables. Again, there are some stark differences on the coefficients depending on O_α. For instance, low O_α values imply that reciprocity r and temporal stability TS, and higher O_α is instead associated to a higher effect of N^E.

5.3 Discussion

Tie strengths are a multifaceted phenomena, and this thesis plays testament to that. Indeed, tie strengths defined in terms of community structures admit several behavioural characterizations, so that *communication intensity yields only partial information* of ties. Indeed, it is striking that the most relevant variables (at least in terms of average MCC performance across different models) are conceptually different among them: from the number of bursty trains N^E to differences in daily behaviour JSD_{diff} to temporal stability TS and activity at specific hours -such as weekend afternoons ($C12$), all provide valuable information on tie strengths.

When it comes to variable interactions, it is noteworthy that different models yield highly different results, highlighting the fact that some of these variables might interact in non-trivial ways. Interestingly, our communication intensity measure showed the strongest interaction with variables of behavioural differences in daily patterns (JSD and JSD_{diff}). This might have to do with the fact that these variables have little correlation to communication intensity, being derived mostly from nodes' overall activity, revealing latent structures in the process.

It is surprising that some variables gain significant prominence for the mean temporal overlap model, yet where not so for the full static overlap model, as it might imply that there is a stronger relationship between changes in local topology and tie-level behaviour than previously thought. In particular, the mean inter event-time $\bar{\tau}$ and temporal stability TS become relevant for this model - both variables being tightly related to the network topology of the tie itself under our model for for temporal overlap. This points to an interesting direction in analysis, particularly for detection of changes in community structures given behavioral data.

For construction of networks from CDR or similar data, we find that we may use a wide variety of weights and retain the Granovetter effect. This is good news, as it may yield a positive overall effect on further studies, adding versatility and more modelling choices. The exact choice of feature for tie strength - number of bursty trains, communication intensity, temporal stability or divergence of daily patters, for instance, might need to be tailored for specific studies. Alternatively, it might be possible and desirable to include multidimensional characterizations of tie strengths.

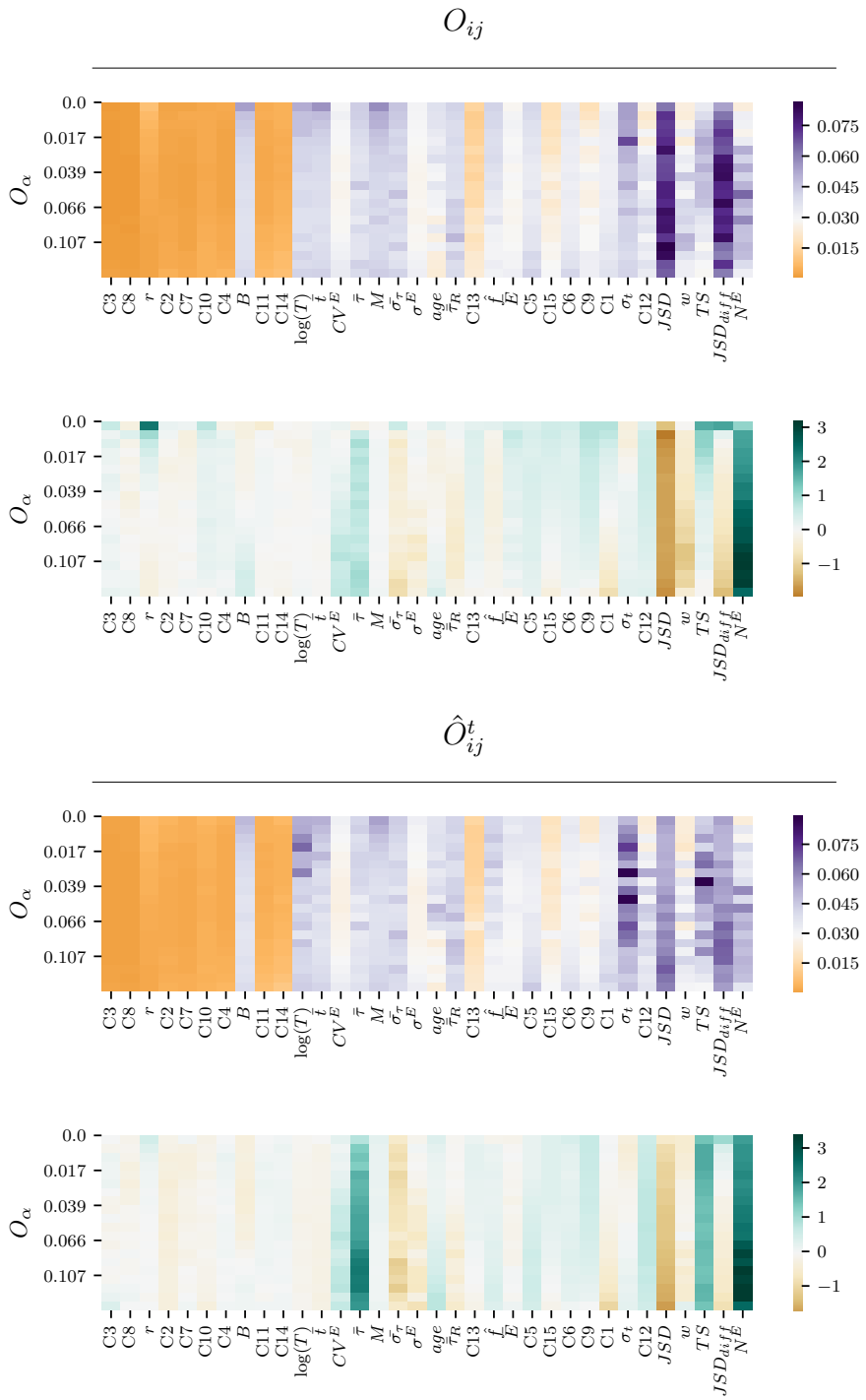


Figure 34: Feature importances for predicting full sets of variables for predicting (*top two*) static overlap O_{ij} and (*bottom two*) mean temporal overlap $\hat{O}^t_{ij}$. For each target variable, the top graph represents feature importances for RF, while the bottom graph represents the coefficients of LR. Since RF and ET behave similarly for these sets of variables, we only show feature importances for RF and LF. For RF purple (orange) values represent features of higher (lower) importance. For LR the color represents the sign, and the intensity the magnitude of the effect.

6 Conclusions

On this thesis we explored different temporal features, most of which were based on previous research, in an attempt to understand their relationship to network structure. We did this under the assumption that community-structures around a tie are coupled with a latent variable of the *strength of a tie*, an elusive concept that is of high importance in social networks. By doing so, we challenged the notion that communication intensity measures - such as the total number of calls, are the optimal choices for quantifying tie strengths. Indeed, we found that although these variables are relevant, they contain only partial information about human communication, which admits much more varied characterizations than intensity-derived statistics. The results of this thesis may be used to make a better-informed decision when it comes to building social networks from communication data using - for example, the number of bursty trains, temporal stability or differences in daily behaviour.

We divided this thesis into two main sections based on different conceptions of time. In the first case, we modelled time linearly, where we focused on how communication events are distributed as a sequence of events. We replicated and derived various measures that focused on different aspects of event-based behaviour, focusing on three main areas: the distribution of inter-call times, bursty processes, and temporal stability. Remarkably, all three areas proved to be fruitful for obtaining measures that could be associated to tie strengths. In particular the idea of counting bursty cascades (N^E) proved to be a useful measure that, in a way, penalizes high-intensity activity in time (bursts). Temporal stability (TS) is another simple idea -the fraction of time where we observe communication out of the whole observation window, that proved to be promising, particularly when measuring overlap in time.

The next characterization of time focused on cycles: activity patterns over days and weeks. We measured how people place calls during the day, and compared the daily activity distribution, revealing behavioural homophily: similar activity patterns are associated to community structures. We also analyzed whether different calling time profiles during the week were more or less associated to network topology, finding the (probably culturally-bound) times when people talked more to other people where we could associate a community structure.

In addition, we introduced two main results related to the measurement of

topological overlap. First, we showed that there was a structural bias in our data: we were more likely to observe certain nodes than others. All in all, we proved that if the probability of observing common neighbors of two nodes is higher than the probability of observing non-common neighbors, then the estimates are biased. Indeed, this does seem to be the case for a majority of links, encouraging the use of non-sampled networks when the data is available, and alerting of the situation in case it is not. Second, we analyzed how topological overlap evolves in time, with results suggesting that many ties that define common neighbors occur at rather distinct moments in time.

6.1 Further research

This thesis was by no means meant to be an exhaustive compilation of behavioral and topological features in communication networks, and there are many new research directions. For instance, we could analyze higher-level communication correlations: how likely are two people to talk to each other after/before talking to other people? We could also add simpler lower-level topological features, such as differences in degrees; add more intensity-based measurements, or focus on the distribution of call lengths, to see whether there are any possible temporal correlations.

Considering the features we did use, there is a need to further analyze some variables. For instance, when dealing with bursty cascades, an analysis of sensitivity to the parameter Δt is necessary, particularly given the importance of N^E in our models. Indeed, we argued that the authors [21] performed a sensitivity analysis and found that the method is robust for a large range of values -and preliminary tests we performed seemed to confirm this-, yet there could be other interesting associations to overlap. There could also be a more thorough examination of the clusterization effect on our weekly profiles, with a different approach probably focusing more on the *best* clusters, and not on capturing overlap variation. In addition, we did not explore many relationships between our variables, which could be a helpful direction for a more thorough modelling of overlap.

For the daily and weekly patterns, there are a myriad of new directions. First, it would be interesting to add more nuance to the differences in distributions, and see whether JSD differs when comparing weekday and weekend activity. We could

also develop a non-biased construction of our variable $JSD_{i\to j}$, and focus on a more theoretical modelling of differences in daily distributions. In addition, finding *why* the differences in daily activity are associated to overlap could be interesting - particularly considering that there could be latent variables such as age. As for weekly patterns, new research could focus on the interactions between different clusters and on more specific network structures that might appear at different times. In other words, we could perform a more thorough analysis to see whether more complex weekly profiles contain more information about overlap (we mentioned Friday nights against workday mornings, for example). A different approach would be to analyze a multiplex network based on different times, and measure overlap on different layers of a network.

We do not provide specific combinations of variables that might be used as stronger indications of tie strength, we only point to different variables that capture the Granovetter effect. For this reason, there could be a more thorough analysis of specific combinations of variables for specific scenarios.

Last, we only performed an exploratory analysis of how overlap varies in time. Indeed, this is a rich area for further research, particularly since we know communication networks to be highly temporal objects, and *time* should be explicitly modelled when dealing, for instance, with larger observation windows. There should be a larger focus on the effect of different parameters (we only tested temporal overlap for $\Delta T = 1$ month), as well as more thorough examination of the resulting time-series of overlap values, not only the mean.

References

[1] Adan, A., Archer, S. N., Hidalgo, M. P., Milia, L. D., Natale, V., and Randler, C. Circadian typology: A comprehensive review. *Chronobiology International 29*, 9 (2012).

[2] Aledavood, T., Lehmann, S., and Saramäki, J. Digital daily cycles of individuals. *Frontiers in Physics 3* (2015), 73.

[3] Aledavood, T., Lehmann, S., and Saramäki, J. Social network differences of chronotypes identified from mobile phone data, 2017.

[4] Aledavood, T., López, E., Roberts, S. G. B., Reed-Tsochas, F., Moro, E., Dunbar, R. I. M., and Saramäki, J. Daily rhythms in mobile telephone communication. *PLOS ONE 10*, 9 (2015).

[5] Aledavood, T., López, E., Roberts, S. G. B., Reed-Tsochas, F., Moro, E., Dunbar, R. I. M., and Saramäki, J. Channel-specific daily patterns in mobile phone communication, 2015.

[6] Barabási, A.-L. The origin of bursts and heavy tails in human dynamics. *Nature 435*, 7039 (may 2005), 207–211.

[7] Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences 101*, 11 (2004).

[8] Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. Network analysis in the social sciences. *Science 323*, 5916 (feb 2009), 892–895.

[9] Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical 41*, 22 (2008), 224015.

[10] Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Mining 10* (2017).

[11] Choudhury, M. D., Mason, W. A., Hofman, J. M., and Watts, D. J. Inferring relevant social networks from interpersonal communication. In *Proceedings of the 19th international conference on World wide web* (2010), ACM Press.

[12] Clauset, A., Shalizi, C. R., and Newman, M. E. J. Power-law distributions in empirical data. *SIAM Review 51*, 4 (2009), 661–703.

[13] Goh, K.-I., and Barabási, A.-L. Burstiness and memory in complex systems. *EPL (Europhysics Letters) 81*, 4 (jan 2008), 48002.

[14] González, M. C., Hidalgo, C. A., and Barabási, A.-L. Understanding individual human mobility patterns. *Nature 453*, 7196 (jun 2008), 779–782.

[15] Granovetter, M. S. The strength of weak ties. *American Journal of Sociology 78*, 6 (1973), 1360–1380.

[16] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

[17] Holme, P., and Saramäki, J. Temporal networks. *Physics Reports 519*, 3 (oct 2012), 97–125.

[18] Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., and Mazaid, M. Opening closed regimes: What was the role of social media during the arab spring? *SSRN Electronic Journal* (2011).

[19] Jo, H.-H. Analytically solvable autocorrelation function for correlated interevent times. 3.

[20] Jo, H.-H., Karsai, M., Kertész, J., and Kaski, K. Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics 14*, 1 (jan 2012), 013055.

[21] Karsai, M., Kaski, K., Barabási, A.-L., and Kertész, J. Universal features of correlated bursty behaviour, 2012.

[22] Karsai, M., Kaski, K., and Kertész, J. Correlated dynamics in egocentric communication networks. *PLoS ONE 7*, 7 (jul 2012), e40612.

[23] Karsai, M., Kivelä, M., Pan, R. K., Kaski, K., Kertész, J., Barabási, A.-L., and Saramäki, J. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E 83*, 2 (2011).

[24] Kim, E.-K., and Jo, H.-H. Measuring burstiness for finite event sequences. *Physical Review E 94*, 3 (Sept. 2016).

[25] Kivelä, M., Pan, R. K., Kaski, K., Kertész, J., Saramäki, J., and Karsai, M. Multiscale analysis of spreading in a large communication network. *Journal of Statistical Mechanics: Theory and Experiment 2012*, 03 (mar 2012), P03005.

[26] Kivelä, M., and Porter, M. A. Estimating interevent time distributions from finite observation periods in communication networks. *Physical Review E 92*, 5 (nov 2015).

[27] Kolaczyk, E. D. *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated, 2009.

[28] Lee, J. H., Kim, I. S., Kim, S. J., Wang, W., and Duffy, J. F. Change in individual chronotype over a lifetime: A retrospective study. *Sleep Medicine Research 2*, 2 (2011).

[29] Miritello, G. *Temporal Patterns of Communication in Social Networks*. Springer International Publishing, 2013.

[30] Miritello, G., Lara, R., Cebrian, M., and Moro, E. Limited communication capacity unveils strategies for human interaction. *Scientific Reports 3*, 1 (jun 2013).

[31] Miritello, G., Lara, R., and Moro, E. Time allocation in social networks: Correlation between social structure and human communication dynamics. In *Understanding Complex Systems*. Springer Berlin Heidelberg, 2013, pp. 175–190.

[32] Miritello, G., Moro, E., and Lara, R. Dynamical strength of social ties in information spreading. *Physical Review E 83*, 4 (apr 2011).

[33] Miritello, G., Moro, E., Lara, R., Martínez-López, R., Belchamber, J., Roberts, S. G., and Dunbar, R. I. Time as a limited resource: Communication strategy in mobile phone networks. *Social Networks 35*, 1 (jan 2013), 89–95.

[34] Murnane, E. L., Abdullah, S., Matthews, M., Kay, M., Kientz, J. A., Choudhury, T., Gay, G., and Cosley, D. Mobile manifestations of alertness. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services* (2016), ACM Press.

[35] Navarro, H., Miritello, G., Canales, A., and Moro, E. Temporal patterns behind the strength of persistent ties. *EPJ Data Science 6*, 1 (dec 2017).

[36] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. Structure and tie strengths in mobile communication networks, 2007.

[37] Panda, S., Hogenesch, J. B., and Kay, S. A. Circadian rhythms from flies to human. *Nature 417*, 6886 (2002).

[38] Park, P. S., Blumenstock, J. E., and Macy, M. W. The strength of long-range ties in population-scale social networks. *Science 362*, 6421 (Dec. 2018), 1410–1413.

[39] Raeder, T., Lizardo, O., Hachen, D., and Chawla, N. V. Predictors of short-term decay of cell phone contacts in a large scale communication network. *CoRR abs/1102.1753* (2011).

[40] Ramos, R., Sassi, R., and Piqueira, J. Self-organized criticality and the predictability of human behavior. *New Ideas in Psychology 29*, 1 (jan 2011), 38–48.

[41] Reka, A., and Barabasi, A. Statistical mechanics of complex networks, 2002.

[42] Sah, P., Méndez, J. D., and Bansal, S. Animal social network repository, 2018.

[43] Saramaki, J., Leicht, E. A., Lopez, E., Roberts, S. G. B., Reed-Tsochas, F., and Dunbar, R. I. M. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences 111*, 3 (jan 2014), 942–947.

[44] Saramäki, J., and Moro, E. From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. *The European Physical Journal B 88*, 6 (2015).

[45] Tsaousis, I. Circadian preferences and personality traits: A meta-analysis. *European Journal of Personality* (2010).

[46] Van Dongen, S. A new cluster algorithm for graphs. *Inf. Syst. 1* (08 2002).

[47] Vicario, M. D., Zollo, F., Caldarelli, G., Scala, A., and Quattrociocchi, W. Mapping social dynamics on facebook: The brexit debate. *Social Networks 50* (July 2017), 6–16.

[48] Watts, D. J., and Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature 393*, 6684 (June 1998), 440–442.

[49] Wuchty, S. What is a social tie? *Proceedings of the National Academy of Sciences 106*, 36 (sep 2009), 15099–15100.

[50] Wuchty, S., and Uzzi, B. Human communication dynamics in digital footsteps: A study of the agreement between self-reported ties and email networks. *PLoS ONE 6*, 11 (nov 2011), e26972.